MINING THE DIGITAL INFORMATION NETWORKS
Mining the Digital Information Networks
Proceedings of the 17th International Conference on Electronic Publishing

Edited by
Niklas Lavesson
*Blekinge Institute of Technology, Sweden*

Peter Linde
*Blekinge Institute of Technology, Sweden*

and

Panayiota Polydoratou
*Alexander Technological Educational Institute of Thessaloniki, Greece*

IOS Press
Amsterdam • Berlin • Tokyo • Washington, DC
Preface

The main theme of the 17th International Conference on Electronic Publishing (ELPUB) concerns different ways to extract and process data from the vast wealth of digital publishing and how to use and reuse this information in innovative social contexts in a sustainable way. We bring together researchers and practitioners to discuss data mining, digital publishing and social networks along with their implications for scholarly communication, information services, e-learning, e-businesses, the cultural heritage sector, and other areas where electronic publishing is imperative.

ELPUB 2013 received 36 paper submissions. The peer review process resulted in the acceptance of 16 papers. From the accepted papers, 8 were submitted as full papers and 8 as extended abstracts. These papers were grouped into sessions based on the following topics: Data Mining and Intelligent Computing, Publishing and Access, and Social Computing and Practices.

James MacGregor and Karen Meijer-Kline from the Public Knowledge Project (Simon Fraser University Library, Canada) lead the pre-conference workshop on June 12. The workshop is entitled “The Future of E-publishing: An Introduction to Open Journal Systems & Open Monograph Press”.

The main program on June 13–14 features two keynotes. Stephan Shakespeare (YouGov, UK) will deliver a keynote entitled “Getting value out of our digital trace: a strategy for unleashing the economic and justice potential of data sharing”. Professor Felix S. Wu (University of California at Davis, USA) will deliver a keynote entitled “Social computing leveraging online social informatics”. ELPUB 2013 also features a panel discussion entitled “Setting research data free – problems and solutions”. The panel consists of the aforementioned keynote speakers as well as Professor David Rosenthal (Stanford University, USA) and Hans Jörgen Marker (Swedish National Data Service, Sweden).

We believe that the topics featured in the program of this year’s ELPUB conference are diverse and exciting. Firstly, we would like to thank members of the ELPUB Executive Committee who, together with the Local Advisory Committee, provided valuable advice and assistance during the entire process of the organization. Secondly, we would like to thank our colleagues in the Program Committee who helped in assuring the quality of the conference throughout the peer reviewing process. Lastly, we acknowledge the Local Organization team for making sure that all efforts materialized into a very interesting scientific event. Thank you all for helping us maintain the quality of ELPUB and deserve the trust of our authors and attendees.

We wish you all a good conference and we say farewell hoping to see you again in Greece for the next installment of the conference in 2014!

Niklas Lavesson, Peter Linde, and Panayiota Polydoratou (editors)
This page intentionally left blank
Organizing Chairs

**General Chair:** Niklas Lavesson – Blekinge Institute of Technology (Sweden)

**Programme Chair:** Panayiota Polydoratou – Alexander Technological Educational Institute of Thessaloniki (Greece)

**Organizing Chair:** Peter Linde – Blekinge Institute of Technology (Sweden)

Executive Committee

John Smith – University of Kent at Canterbury (UK)
Peter Linde – Blekinge Institute of Technology (Sweden)
Karel Jezek – University of West Bohemia in Pilsen (Czech Republic)
Ana Alice Baptista – University of Minho (Portugal)
Sely Costa – University of Brasilia (Brazil)
Jan Engelen – Catholic University of Leuven (Belgium)
Micheál Mac An Airchinnigh – Trinity College Dublin (Ireland)
Bob Martens – Vienna University of Technology (Austria)
Leslie Chan – University of Toronto (Canada)
Susanna Mormati – CILEA (Italy)
Turid Hedlund – Swedish School of Economics and BA, Helsinki (Finland)
Yasar Tonta – Hacettepe University (Turkey)
Alexander Technological Educational Institute of Thessaloniki (Greece)
Milena Dobreva, University of Malta

Local Organization Team

Maria Engberg – Blekinge Institute of Technology (Sweden)
Bo Helgeson – Blekinge Institute of Technology (Sweden)
Henric Johnson – Blekinge Institute of Technology (Sweden)
Michaela Linge – Blekinge Institute of Technology (Sweden)
Eva Norling – Blekinge Institute of Technology (Sweden)
Ewa Ośkowska – Blekinge Institute of Technology (Sweden)
Programme Committee

Alvim, Luísa – University of Évora (Portugal)
Ayris, Paul – University College of London (UK)
Baptista, Ana Alice – University of Minho (Portugal)
Björnshauge, Lars – SPARC Europe (Denmark)
Bolter, Jay David – Georgia Institute of Technology (USA)
Borges, Maria Manuel – University of Coimbra (Portugal)
Chan, Leslie – University of Toronto Scarborough (Canada)
Coleman, Ross – University of Sydney (Australia)
Delgado, Jaime – Universitat Politècnica de Catalunya (Spain)
Engberg, Maria – Blekinge Institute of Technology (Sweden)
Engelen, Jan – Katholieke Universiteit Leuven (Belgium)
Fitzpatrick, Kathleen – Modern Language Association (USA)
Gabriel, David – University of Porto (Portugal)
Garoufallou, Emmanouel – Technological Educational Institute of Thessaloniki (Greece)
Gathegi, John – University of South Florida (USA)
Griffin Speer, Julie – Virginia Tech (USA)
Hagerlid, Jan – Consultant (Sweden)
Hedlund, Turid – Hanken School of Economics (Finland)
Jamali, Hamid – Kharazmi University (IRAN)
Jezek, Karel – University of West Bohemia in Pilsen (Czech Republic)
Johnson, Ian M. – The Robert Gordon University (Scotland)
Koprinska, Irena – University of Sydney (Australia)
Kurbanoglu, Serap – Hacettepe University (Turkey)
Lavesson, Niklas – Blekinge Institute of Technology (Sweden)
Linde, Peter – Blekinge Institute of Technology (Sweden)
Mac An Airdhinnigh, Micheal – Trinity College Dublin (Ireland)
Martens, Bob – Vienna University of Technology (Austria)
Martzoukou, Konstantina – University of Malta (Malta)
Mele, Salvatore – CERN (Switzerland)
Moreira, Fernando – Portucalense University (Portugal)
Mornati, Susanna – CILEA (Italy)
Nisheva-Pavlova, Maria – Sofia University (Bulgaria)
Polydoratou, Panayiota – ATEI of Thessaloniki (Greece)
Pryor, Graham – Digital Curation Centre (UK)
Quaresma, Paulo – University of Évora (Portugal)
Rauber, Andreas – Vienna University of Technology (Austria)
Sachini, Evi – National Documentation Centre (Greece)
Savenije, Bas – Utrecht University Library (The Netherlands)
Schranz, Markus – Pressetext Austria (Austria)
Schimmer, Ralf – Max Planck Digital Library (Germany)
Serrano, Anabela – Politechnical Institute of Porto (Portugal)
Slavic, Aída – UDCC (UK)
von de Sompel, Herbert – Los Alamos National Laboratory (USA)
Sousa Pinto, Joaquim – University of Aveiro (Portugal)
Strufe, Thorsten – TU Darmstadt (Germany)
Svensson, Aina – Uppsala University (Sweden)
Tenopir, Carol – University of Tennessee (USA)
de Waard, Anita – Elsevier Labs and Utrecht Institute of Linguistics (Netherlands)
Warner, Simeon – Cornell University (USA)
Virkus, Sirje – Tallinn University (Estonia)
Wu, Felix S. – University of California at Davis (USA)
This page intentionally left blank
## Contents

Preface \hspace{1cm} v  
*Niklas Lavesson, Peter Linde and Panayiota Polydoratou*

Organizing Chairs, Executive Committee, Local Organization Team, Programme Committee \hspace{1cm} vii

### Full Research Articles

#### Data Mining and Intelligent Computing

- The Shattered Document Approach to Adaptive Hypertext. Design and Evaluation  
  *Mário Amado Alves*  
  Page 3
- 3D Gesture-Based Exploration and Search in Document Collections  
  *Angélica de Antonio, Cristian Moral, Daniel Klepel and Martín J. Abente*  
  Page 13
- Cover Sheets Considered Harmful  
  *Emma L. Tonkin, Stephanie Taylor and Gregory J.L. Tourte*  
  Page 23

#### Publishing and Access

- Me4DCAP V0.1: A Method for the Development of Dublin Core Application Profiles  
  *Mariana Curado Malta and Ana Alice Baptista*  
  Page 33

#### Social Computing and Practices

- COMFY – A Conference Management Framework  
  *Christian Caldera, René Berndt and Dieter W. Fellner*  
  Page 45

### Full Professional Articles

#### Data Mining and Intelligent Computing

- Deduplication of Metadata Harvested from Open Archives Initiative Repositories  
  *Piotr Wendykier*  
  Page 57

#### Publishing and Access

- Facilitating Access and Reuse of Research Materials: The Case of the European Library  
  *Nuno Freire*  
  Page 67
- PlosOpenR Exploring FP7 Funded PLOS Publications  
  *Najko Jahn, Martin Fenner and Jochen Schirrwagen*  
  Page 77
Extended Abstracts

The extended abstracts are available in full text online from: http://elpub.scix.net/cgi-bin/works/Home

Data Mining and Intelligent Computing

Making Sense of a Flat List of Terms into Linked Open Data SKOS Vocabularies
Helder Noel Firmino and Ana Alice Baptista 89

Publishing and Access

Does Openness and Open Access Policy Relate to the Success of Universities
Pekka Olsbo 90

Usage Data of an Open Access e-Journal in a Digital Repository
Teresa S. Welsh and Stacy Creel 91

Publisher Interest Towards a Role for Journals in Data-Sharing: The Findings of the JoRD Project
Marianne Bamkin, Jane Smith, Paul Sturges, Azhar Hussein and Bill Hubbard 92

Opening Up Digital Publications – Suggestions for Making Humanities Data Available for Research
Henriette Roued-Cunliffe 93

Repositories Recreated – Working Towards Improved Interoperability and Integration by a Co-Operative Approach in Sweden
Stefan Andersson and Aina Svensson 94

Towards an e-Publishing Library Service in Humanities and Social Sciences: A Feasibility Study
Rosa Di Cesare, Silvia Gianmini and Daniela Luzi 95

Social Computing and Practices

Making Social Interactions Accessible in Online Social Networks
Fredrik Erlandsson, Roosbeh Nia, Henric Johnson and Felix Wu 97

Subject Index 99
Author Index 101
Full Research Articles
This page intentionally left blank
The Shattered Document Approach to Adaptive Hypertext. Design and Evaluation

Mário AMADO ALVES a,1
aLIAAD-INESC TEC. amado.alves@gmail.com

Abstract. We address the problem of improving, automatically, the usability of a large online document. We propose an adaptive hypertext approach, based on splitting the document into components smaller than the page or screen, called noogramicles, and creating each page as a new assemblage of noogramicles each time it is accessed. The adaptation comes from learning the navigation patterns of the users (authors and readers), and is manifested in the assemblage of pages. We test this model across a number of configurations, including chance and non-adaptive systems. We evaluate our model through simulation. We have designed a simulator based on established findings about the behaviour of hypertext users. We have realised a quantitative evaluation based on hypertext usability measures adapted to the problem: session size, session cost.

Keywords. H.3.3 Information Search and Retrieval / Information filtering, Retrieval models, Search process, H.5.1 Multimedia Information Systems / Evaluation/methodology, H.5.4 Hypertext/Hypermedia / Theory, I.2.0 General / Cognitive simulation, I.2.6 Learning / Connectionism and neural nets, Algorithms, Design, Experimentation, Theory, adaptive hypertext, spreading activation

1. Introduction

We study the large online document, and how its utilization might be improved by means of adaptive hypertext features. Large means an extent such that the document cannot be seen all at once. In other words: large = (much) larger than a screenful.

The overall process of hypertextualization and hypertext adaptation is depicted in figure 1. The Document state represents a conventional document, or else a poorly structured para-document like a set of forum posts, a wiki, etc. The hypertextualization step consists in casting this content into an Hypertextual form that can be adapted. The aim of adaptation is to have this form evolve onto a Better hypertext, by learning from the navigation patterns of the users (or some other adaptive input).

A classical example of adaptive hypertext at the service of improved utilization of a large online document is the Knowledge Sea II system [1], applied to a manual of C, in a programming language learning environment.

1This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.
In our work, the hypertextualization step entails splitting—shattering—the Document into its atomic constituent parts, which we call noogramicles (small representations of knowledge). We detail this model later. We use as adaptative input the choices of the users. The adaptative output consists in newly created pages, assembled from the most connected noogramicles; therefore we call this process renoding—paraphrasing the classical method of relinking [5].

1.1. Operational definition of hypertext

There are various kinds of hypertext, cf. [9], [7] and references therein. A consensual definition is missing at large. From the examination of the various kinds of hypertext, we have found the following invariants, the set of which constitutes our operational definition:

- hypertext is an interface to interconnected items
- the items are of information, textual or pictorial
- the interface lets the user follow any connection
- the interface records the connections followed, and lets the user relive them at will; in particular, the interface provides a back button

1.2. Article organization

On section 2 we describe the shattered document approach, which comprises a document model (section 2.1) and an adaptation model (section 2.2). On section 3 we describe our experimental design, namely the simulator (section 3.1) and the quality measures used to evaluate the configurations (section 3.2). On section 4 we describe a selection of the configurations experimented with and their results. We conclude on section 5.

For space reasons, in this article we had to leave out a number of items, notably the rationalia for hypertextualization and renoding, the description of the method of spreading activation, and the description and results of a large number of experimental configurations pertaining to random users and alternative document structures. Such items are fully described elsewhere [3,2].
Figure 2. Standard hypertextualization of the sequential structure.

Figure 3. Model of the same document in figure 2 but with the shattered document approach and the two types of connection Next (N) and Child (C).

2. The Shattered Document Model

2.1. Document Model

The Shattered Documents approach prescribes that documents be taken apart into their smallest constituents of meaning, or noogramicles. Naturally the noogramicles must connect with each other, in order to create, ultimately, a navigatable network, or hypertext.

We look mostly at hypertext documents created from traditional documents, for example the hypertext version of the *Ada Reference Manual* or ARM [4], which we use for experimentation. We have observed three dimensions in the traditional structure of documents—sequence, hierarchy, cross-reference. We transcribe these dimensions into types of connection in the network.

Figure 2 depicts the standard hypertextual edition of the ARM. The representation, albeit stylised, honors the actual data, in the numbering and size of the sections, and the links thereof. Each section is a node, or page—an integral HTML file. The pages are connected by *Next* and *Previous* links. Four sequent nodes in this sequence are represented. The links are symbolized by the arrows in figure 2, and designed on the interface as buttons located at the top and at the bottom of each page.

Figure 3 shows how the same part of the ARM in 2 is modelled with the shattered document approach and the two types of connection Next and Child. Figure 4 shows a page made up of the first five noogramicles in the model: the constituent noogramicles of this page are selected using spreading activation from the first one in a manner detailed later.
To consolidate: a document is represented as a graph, or network data model of noogramicles interrelated by directed connections of three types—Next, Child, Refer—as follows.

**Next** represents the linear order of paragraphs. Example: from a paragraph to its immediate successor. Note that, by the extended paragraph definition, Next also connects from a section heading to the first classic paragraph of the section, and from the last classic paragraph of a section to the heading of the next section.

**Child** represents the immediate subordinate relationship between paragraphs. Examples: from a section to each of its subsections; from a paragraph introducing an enumeration (e.g. a bulleted list) to each item of the enumeration; possibly, even from a terminal section (i.e. a section without subsections) to each of its paragraphs.

**Refer** represents other reference relationships. Examples: from a paragraph to a footnote; from a paragraph to another paragraph or section (e.g. the so-called cross references, and see also references); from a index entry to its target paragraph or section; from a TOC entry to the corresponding section.

In the Child and Refer relationships, a section is represented by its first paragraph, normally a heading. Original references anchored in sub-paragraph units (e.g. words) are represented as references anchored on the paragraph as a whole.

As we are targeting an adaptive system, we need a way to represent the corresponding information. The adaptative part of our model is compound of two main items:

**Pages.** What the reader sees. A page is assembled from a small number of noogramicles, in a manner detailed later. Naturally the user may navigate to another page. Pages are the adaptive output of shattered documents.

**Travels.** The navigation steps that users (authors and readers) make in the document. Each travel is recorded, and used in adapting the construction of pages. Travels are the adaptive input of shattered documents.

The adaptative process integrates the two items, by assembling pages based on travel information. The main idea is to select the noogramicles that are most connected to the current one.

So, we must add the connection type Travel to the trio Next, Child, Refer already explained. Therefore, so far our document model is a network of noogramicles with four types of connection: Child, Next, Refer, Travel.

In the current experimental configurations, we interpret each of Child, Next, Refer, as Travel. That is, we unify all types into one. This reinterpretation of the traditional document structure connection types Child, Next, Refer allows us to solve the cold start problem, and simplifies immensely the process of spreading activation. This magic step is justified mainly because, if you look at it, the connections Child, Next, Refer are indeed
the travels that the author intended the reader to make in the first place. Next is directly so. Child, Refer are contigently so—they are the paths laid out by the author for the reader to cross, wanting. Or, Child, Refer carry a rhetorical value—which amounts to the same effect (a contigent choice by the reader to follow).

So, a shattered document is represented as a graph, or network datamodel of noogramicles interrelated by directed connections representing travels: the paths either actually taken by users or else suggested by the author.

2.2. Adaptation Model

Our main architectural ideas are as follow. The system is an interface into a large document. The interface unit is the page. Pages are accessed one at a time, normally. The document as a whole is partitioned, shattered, into fragments smaller than the page, called noogramicles. Each page is an assemblage of noogramicles.

Each noogramicle has two renderings, or views: expanded, contracted. The expanded view is the noogramicle itself, normally. The contracted view is a clear label of the noogramicle. Occasionally, the label equates the noogramicle, i.e. their contracted and expanded views are formally identical.

Figure 4 exemplifies one page in our design. The noogramicle on the top is central, and represents the page for certain purposes, explained below. On this design, the real estate on a page is divided approximately equally between expanded and contracted noogramicles. The higher-ranking noogramicles are expanded.

Navigation, or travelling, on this design, is effected by recentring on a noogramicle, normally by clicking on it.

The main adaptive input of our system consists of the choices, or travels, made by the users. These travels are memorized in the computer as connections between the respective noogramicles. The target point of the connection is the noogramicle clicked on. The source point of the connection is the central noogramicle of the current page, normally. This configuration is called central-to-central. Other configurations are possible and were tested, e.g. all-to-central, but are not reported in this article.

This graph model is then explored (in the computer) by means of spreading activation [3], to find the noogramicles more related to the central one (which has been chosen by the user). Such noogramicles then form the new page—the adaptive output. Namely, energy is spread from the central noogramicle, and the $n$ most active noogramicles are selected for the new page, where $n = 10$ in the configurations using this method (exposed later).

3. Experimental design

3.1. The Simulator

We have constructed a simulator of user utilization of an adaptive hypertext system, to experiment and compare different configurations of the adaptive system, automatically. The design bases of the simulator include the concept of oracle, and the Smart User Assumption.

The simulator creates sessions of utilization of a hypertext. The simulator comprises two components: the Oracle Model and the User Model. Each session is a user’s quest for
the noogramicle that will respond to their reference question, or information need. Such noogramicle is called the oracle of the session. The simulator draws a random oracle for each session. The random distribution of oracles is called the Oracle Model.

The simulated user, or User Model consists in a function Choice which selects a page item. It acts as the user clicking on a hyperlink to follow. The Smart User Assumption asserts that a human user chooses the right link if the link label is clear about the distal content. For the case when the oracle is only one click away, the label clearly identifies the oracle, and therefore the user selects the item easily. When the oracle is further away than one click, the intelligence or intuition of the user takes place to select the item most likely to lead to the goal. The results in [5] and [10] are interpretable as support for this assumption.

Given these premises, the Choice function can be designed as a choice of the page item most likely to lead to the oracle. In our implementation this is done via spreading activation from the atom, until a page item $X$ is reached, such that $X$ has not been seen or visited before in the same session. The most active item represents the item most connected to the oracle, globally.

The session ends upon reaching the oracle (successful session), or else a maximum session size (number of pages) has been breached (unsuccessful session).

3.2. Quality measures

We introduce the hypertext usability or quality measure of session cost as a refinement of the well established measure of session size = number of pages = number of clicks - 1. Session cost is a combination of session size with a few extra factors of cognitive effort associated with poorer or null navigational aids, longer pages (requiring scroll), and a large quantity of links (requiring more examination). This is necessary because we want to compare across different configurations, and, on the limit, a configuration of only one large page containing the entire document has always the best possible session size of 1—thus unfairly winning the competition even before the start!

It sounds natural, practical, reasonable, to extend the unit of session size, the page, to this new unit of session cost. Let us call it the session cost point, or just point. We formalize session cost as the sum of the cognitive effort terms involved. We have identified the extra cognitive factors described below. For each factor, we introduce a formula for its contribution to the cognitive cost. These factors affect the page. Naturally, the session cost is the summation of the page costs.

We quantify cognitive effort in points directly related to the session size component. We establish a basis: for a configuration of no scroll and a fixed, low page size (quantity of page items, or links) of at most 10, we equate session cost with session size, and therefore define a fixed page cost value of 1 point. From this basis we define page cost for non-trivial pages as follows.

Scroll cost. The cognitive cost associated with pages longer than the screen is related to that length [11].

\[
\text{Scroll Cost} = \left\lfloor \frac{p}{s} \right\rfloor \tag{1}
\]

where $p$ is the page length, and $s$ the screen size, measured in the same units, e.g. characters.
Table 1. Principal statistics compared

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Success</th>
<th>$\text{Size}_0$</th>
<th>$\text{Size}_1$</th>
<th>$\text{S-Gain}$</th>
<th>$\text{Cost}_0$</th>
<th>$\text{Cost}_1$</th>
<th>$\text{C-Gain}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shattered Document</td>
<td>0.94</td>
<td>5.62</td>
<td>3.48</td>
<td>★1.62</td>
<td>5.62</td>
<td>3.48</td>
<td>★1.62</td>
</tr>
<tr>
<td>Shattered Random Document</td>
<td>0.98</td>
<td>4.10</td>
<td>★3.40</td>
<td>1.20</td>
<td>4.00</td>
<td>★3.38</td>
<td>1.18</td>
</tr>
<tr>
<td>Structural</td>
<td>1.00</td>
<td>3.48</td>
<td>3.48</td>
<td>1.00</td>
<td>15.20</td>
<td>15.20</td>
<td>1.00</td>
</tr>
<tr>
<td>Markov Chains</td>
<td>0.84</td>
<td>5.84</td>
<td>4.22</td>
<td>1.38</td>
<td>5.84</td>
<td>4.22</td>
<td>1.38</td>
</tr>
<tr>
<td>Shattered Document with Random</td>
<td>0.02</td>
<td>6.90</td>
<td>6.98</td>
<td>0.99</td>
<td>6.90</td>
<td>6.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Total</td>
<td>3.77</td>
<td>25.94</td>
<td>21.56</td>
<td>6.20</td>
<td>37.56</td>
<td>33.26</td>
<td>6.17</td>
</tr>
<tr>
<td>Average</td>
<td>0.75</td>
<td>5.19</td>
<td>4.31</td>
<td>1.24</td>
<td>7.52</td>
<td>6.65</td>
<td>1.23</td>
</tr>
</tbody>
</table>

**Choice cost.** If the number $n$ of links exceeds the magical number 7 minus or plus 2 (cf. [8]), then an extra cognitive effort is imposed upon the user, proportionate to the number of links. From our basis of a fixed cost of 1 point for 10 or less items (close enough to the magical number of $7 + 2 = 9$), we derive:

$$\text{Choice\_Cost} = \max(1, n/10)$$  \hspace{1cm} (2)

3.3. Result statistics and configurations

We computed a number of statistics to evaluate the results of size and cost: moving average over the whole sequence of sessions, with a window of 50 sessions; average per oracle.

The configurations under experiment are described together with their results in the next section.

4. Experimental results

4.1. Overview of results

Table 1 compiles the principal results for each configuration. The minimum and maximum values in each column except $\text{Success\_rate}$ are emphasized. The best final and gain scores in their columns are ★starred. The statistics in the table are a transposition of selected statistics of the involved configurations, mapped as follows:

**Success** = Micro-average of the success rate of sessions

**Size_0** = Average of first session size
Figure 5. Results of configuration Shattered Document

<table>
<thead>
<tr>
<th>Success_Rate</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size_0</td>
<td>5.62</td>
</tr>
<tr>
<td>Size_1</td>
<td>3.48</td>
</tr>
<tr>
<td>Size_gain</td>
<td>1.62</td>
</tr>
<tr>
<td>Cost_0</td>
<td>5.62</td>
</tr>
<tr>
<td>Cost_1</td>
<td>3.48</td>
</tr>
<tr>
<td>Cost_gain</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Size_1 = Average of last session size

S-Gain = Micro-average of the gain in size = \( \frac{\text{Size}_0 - \text{Size}_1}{\text{Size}_1} \)

Cost_0 = Average of first session cost

Cost_1 = Average of last session cost

C-Gain = Micro-average of the gain in cost = \( \frac{\text{Cost}_0 - \text{Cost}_1}{\text{Cost}_1} \)

We remind that our statistics of gain represent a betterment of the results, and does not denote numerical increase of the quality measures. In fact, gain statistics correspond to a numerical decrease of the quality measures, because, for these measures, less is better. In rigour, as numbers, they are inverse quality measures. So, the gain statistics invert the numerical relation to present a positive value, i.e. a more is better value.

In general, we locate the most important statistics to the right of the tables. And to the bottom in individual configuration tables. Ultimately, we look at Cost_gain. A configuration with greater Cost_gain—all other things being equal—is the winner. But, of course, it is never the case that all other things are equal, hence the need to have the remaining data reported and analysed as well, for a correct interpretation of the results.

4.2. Shattered Document

See figure 5 for the quantitative results of this configuration, and a graphical depiction of the evolution of session cost thereof.

This configuration is our shattered document design of adaptive hypertext, with the original structure of the legacy document, namely a version of the ARM cut down to 1000 noogramicles. This 1000-node structure is a tree of five levels (including root), with an average fanout of 17.85, and an average distance from the root of 3.46—which corresponds to an expected average session size, in the non-adapted version, of 4.46

We observe that the evolution is positive, with a final size lower than that expected value of 4.46 for the non-adapted version. This result shows that adaptation does improve the utilization of a hypertext document.

We observe that the evolution happens quickly, in the first circa 200 sessions.

Note that the session cost equals the session size in this configuration, because there is no extra cognitive effort (cf. above). There is no scroll (each page fits in one screen) and the number of choices per page is fixed at ten (the magical number, cf. above).

We observe that the starting size is greater than the expected average for the non-adapted document. Probably this is due to the small size of the shattered document pages.
Figure 6. Results of configuration Shattered Random Document

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success_Rate</td>
<td>0.98</td>
</tr>
<tr>
<td>Size_0</td>
<td>4.10</td>
</tr>
<tr>
<td>Size_1</td>
<td>3.40</td>
</tr>
<tr>
<td>Size_gain</td>
<td>1.20</td>
</tr>
<tr>
<td>Cost_0</td>
<td>4.00</td>
</tr>
<tr>
<td>Cost_1</td>
<td>3.38</td>
</tr>
<tr>
<td>Cost_gain</td>
<td>1.18</td>
</tr>
</tbody>
</table>

compared to the original. Whereas the original pages each contain links to all its descendants nodes, a shattered document page has a fixed number of total items of 10, which is less that the average fanout of 17.85 of the original; therefore, there are oracles that, in the original document are just one node away on a 20-item page (say), but in the shattered document might require an extra, intermediary 10-item page to be visited.

4.3. Shattered Random Document

See figure 6 for the quantitative results of this configuration, and a graphical depiction of the evolution of session cost thereof.

This configuration is like Shattered Document, except the initial state of the document is a random connection of each noogramicle to ten others.

We observe that the rate of evolution is small, but the absolute values are very good. The starting session size is already better that the average of 4.46 for the original structure. The final size is similar to Shattered Document.

This result is interesting because it indicates that a legacy, authored structure might be irrelevant for adaptation.

We observe that the session size results differ slightly from the session cost results. We currently have no definitive explanation for this.

4.4. Markov Chains

See figure 7 for the quantitative results of this configuration, and a graphical depiction of the evolution of session cost thereof.

This configuration is like Shattered Document, but using a standard technique for learning the user patterns, namely Markov Chains (first order) [6], instead of our nominal technique of memorizing all user travels individually.

We observe that the results are positive, but not as good as Shattered Document (or Shattered Random Document).

5. Conclusions

The results show that adaptation improves utilization. This was the main result of our work.
The good values for a random document were a surprise, and warrant more investigation. The simulator was also a by-product of note of this work. It allows to experiment with many different configurations, as opposed to experimentation with human users, which is much more costly. Albeit based on sensible and established rules of user behaviour, the simulator still lacks explicit validation as a reliable surrogate of human users. These are possible avenues for future work.

Finally, let us highlight the concept of renoding as a contribution of this work to adaptive hypertext thinking.

References


3D gesture-based exploration and search in document collections

Angélica de Antonio, Cristian Moral, Daniel Klepel, Martín J. Abente
Facultad de Informática, Campus de Montegancedo, Univ. Politécnica de Madrid
{angelica, cmoral, dklepel}@fi.upm.es, mabente@alumnos.upm.es

Abstract. This paper describes an approach towards the interaction with 3D representations of large document collections. The goal was to provide the user with a highly dynamic environment in which even the very mapping strategy to position documents in space can be adjusted by the user depending on the specific task at hand, on his preferences, or on the context. A modification to the FDP algorithm is proposed, as well as a new gesture-based interaction paradigm in which the user can explore and search information in the collection just by simple hand movements. An experimental user evaluation was carried out to investigate the impact of the proposed approach on the precision of the mental model built by users through exploration, on the effectiveness in information search tasks, and on the general user satisfaction and perception of utility.

Keywords. 3D visualization and navigation, 3D interaction, Information Retrieval, Force-Directed Placement, Clustering, k-means

Introduction

Exploration and search of information contained in large document collections is increasingly a need, but many times also a problem. In fact, Internet offers such a huge quantity of information that it is in practice impossible to explore it exhaustively so as to fill a specific information need. In the last 30 years, many techniques have been proposed to visualize document collections. Some of them were developed for 3D virtual environments, as the third dimension increases the information that can be shown to the user. The application of one or another technique often depends on the type of documents to be visualized. For structured documents, specific visualizations highlight the structure and relations between data [1] [2], while for non-structured documents only the content can be analyzed in order to infer as much information as possible. In the last case, visualization techniques are more general and sophisticated, and then more complex [3].

This paper focuses on 3D visualization of non-structured document collections. The visual metaphor chosen, a sphere, is a simple and generic one in order to avoid overloading the visualization. Then, a collection of documents is virtually represented as a cloud of spheres. But visualization is definitely not enough to satisfy the user needs; interaction with the 3D representation is the key aspect to facilitate the user tasks.

The remainder of this document is divided into 4 sections. In Section 1 we detail the general process that must be followed to visualize a document collection in a 3D virtual environment. Section 2 deals with the interaction techniques we have proposed to
allow the user exploring and extracting information from the 3D virtual environment. To evaluate the benefits and drawbacks of our proposal, we carried out an experiment, whose details and results are presented in Section 3. Finally, Section 4 reports the conclusions we have extracted from our work and proposes future research.

1. 3D Visualization of document collections

Based on the model described in [4], the information visualization pipeline can be split into four major stages. The first one consists in gathering, cleaning and preprocessing the documents, aiming at extracting from data attributes of interest that contain relevant information. Selecting these attributes requires a deep understanding of the task that should be carried on with the data and the nature of the data itself. The second stage consists in mapping visual characteristics to each of these attributes. The most common visual characteristics are color, shape, spatial position and movement [5], and they must be selected carefully in order to exploit the natural perception abilities of the users and considering the nature of each attribute [6]. The third stage consists in the generation of a visual representation of the scene making an efficient use of the screen space while keeping the cognitive load as low as possible. The fourth step deals with the interaction between the users and the virtual representation. Interaction is a fundamental issue regarding the design of any information visualization tool. Interaction mechanisms must be designed in order to assist the user’s tasks, and they can act upon any of the previous stages including: select and manipulate a dataset, navigate through the environment and perform actions affecting the data or the system itself.

1.1. Document Preprocessing

In our system, preprocess aims at determining the similarity between pairwise documents. Many complementary treatments have been proposed to undertake this task: normalization, stopwords lists, stemming [7], Part-Of-Speech Tagging [8], Named Entity Recognition [9], and many others. Which of them to use depends on the efficiency and precision required in the retrieval of information. In our case, we look for a dynamic and interactive, and hence efficient, system, even at the expense of precision.

Our preprocessing phase consists of two main parts: the lexical analysis of each document, and clustering the documents into semantic groups based on the distance measure between them. The first part allows us to translate documents into data structures treatable by computers. The process starts with the tokenization of the documents, followed by a normalization of the obtained tokens, for example by clearing accents, punctuation, email addresses, URLs and numbers, among others. All filters are applied to individual tokens, as we use a bag-of-words model, where context and order are not taken into account. Even if this causes the loss of some semantic information, the treatment is much easier and faster. Once tokens are "clean", we proceed to delete the meaningless ones (included in a stopword list), like articles, prepositions, determinants and so on. At the end, the noise, and the initial vocabulary, are reduced, fastening and the process and increasing its precision.

Before proceeding to the second step of preprocessing, we have to represent every document, and its filtered tokens, in a computational way. We have decided to use
multidimensional vectors where every dimension, also called feature, corresponds to a token, and its content is the frequency of the token in the document. This way of representing documents is known as Vector Space Model [10], allows to easily figure out the so-called cosine similarity between two documents by computing the inner product of their normalized unit vector representations. This measure of similitude, calculated for every pairwise documents, is statistical in nature as it only reflects the proportion of tokens that both documents have in common, independently of their semantic meaning. Therefore, it will not reflect faithfully if they share the same topic, but only the rate of words they share. Even if this may seem a drawback, our objective is not to obtain very precise similitude values, but to obtain good enough ones in the shortest possible time.

After this, the system assigns each document to a thematic group by means of a clustering algorithm based on the similarity between documents. In this work the k-means algorithm [11] is chosen, as it is probably the simplest clustering algorithm available and it matches our needs: the number of clusters is dynamically definable by the user and it is fast enough for small sized datasets, allowing the reconfiguration of the clusters in realtime. On the contrary, k-means is not a deterministic algorithm because of the random selection of the initial seeds, and then the algorithm doesn’t always converge to the same clusters [12]. This is not acceptable for an interactive system as the user could get confused. That’s why we have slightly modified the classical k-means algorithm, by executing it \( n \) times, each of them with a different random combination of initial seeds, and choosing the ones that, after executing k-means, minimize the sum of the distances between any document and the centroid of the group to which it belongs.

Finally, in this step we also extract a representative keyword for each document, by retrieving from every vector the feature with the highest value, which represents the meaningful token most frequent in the document. Again, this process is not very precise, but instead it is really simple and fast, which is our goal.

1.2. Assignment of Visual Characteristics

For this work, three attributes of interest were considered: the similarity between documents, the thematic category of each document and the keywords extracted. Taking into account the considerations mentioned in [6], three different visual characteristics were chosen in order to visually represent these attributes.

First, the collection is represented by a tridimensional cloud of spheres, each of them representing a single document. Similarity between documents is visually represented by the spatial distance between the spheres. So, the more the documents are thematically similar, the closer are the spheres. The thematic category is represented by the hue component of color, so that the same category documents have the same color. The hue values of each document are updated every time the number of clusters changes. For the visualization of the keywords, a single label is imposed on the sphere.

In order to obtain a spatial position for every document, the dimensions of the equivalent vectors must be reduced to three, but trying to keep as many similarity information between the documents as possible. In this work the Force-Directed Placement (FDP) technique [13] was used; it simulates attraction and repulsion forces among the documents depending on their similarity measures. Although an FDP approach has big scalability issues, it also has many desirable properties for our goals: good quality of the resulting layout, iterative and realtime positioning process, and ability to be extended by including other factors in the positioning process, as it will be explained in Section 2.2.
2. Interacting with 3D document collections

Besides the visualization of the documents in the 3D environment, it is essential to provide the user with some techniques to navigate and explore the environment. The classical interaction approach is guided by Shneiderman’s famous mantra: Overview first, Zoom, Filter and Details-on-Demand [14], which describes the natural way in which users explore the information.

General interaction mechanisms can be categorized into navigation, manipulation and system control. Navigation mechanisms allow users to explore the information across the different levels of abstraction. At the overview level, it aids users to identify the potential interest areas of information, as done by the FishEye technique [15]. Manipulation mechanisms allow users to effectively interact with the elements of the virtual environment. The most common examples are selecting elements and manipulating them spatially within the environment. Finally, control system techniques deal with the modification of the visual mapping of the data or of the state of the system. An example is the removal of a set of elements from the environment or the saving of the current state of the system. These mechanisms must be designed taking into account the user’s needs together with the tasks and the specific context of application. Here we present two innovative interaction techniques allowing to manipulate the visual mapping, by improving the identification of thematic groups and the visual perception of document similarities.

2.1. Modifying the Number of Document Groups

The first proposed manipulation mechanism allows users to interactively modify the number of potential thematic groups (see Figure 1.a).

The interest of this technique is based on the assumption that every user has different cognitive and perceptual capabilities, and different preferences. Some users might want to split the document collection into as many specific groups as possible, while others might prefer less and more general groups. The need for higher or lower subdivision even depends on the user task (e.g. building a mental model of the collection vs. searching for a specific document) or on the phase in the pursuit of a certain goal (e.g. filter irrelevant information vs. select the best possible source to search a specific information). Our hypothesis is that combining this mechanism with some filtering mechanisms for documents can greatly increase the flexibility of the system while exploring or searching.
2.2. Adjustment of Attraction Forces

This mechanism allows users to modify the relative position of the documents. Although the spatial positioning reflects by itself the similarity between documents, we think the visual clarity of the representation can be improved. In this sense, we propose a modification to the original FDP algorithm, called FDP with Force Control (FDPFC) algorithm, where two new attraction forces, dynamically modifiable by the users, are added.

The first force modifier aims at enhancing the overview visualization. To achieve this, it modifies the spatial position of spheres so documents that belong to the same cluster get closer and are separated from the other clusters (see Figure 1.b). This modification results into a more compact and lean clusters visualization from the overview perspective. Technically, this effect is generated by adding a specific multiplier to the force that controls the attraction between any pairwise documents, and it is applied if and only if both documents belong to the same cluster.

The second force modifier aims at enhancing the inter-document similarity perception, both in the overview and the detail levels. This is accomplished by modifying the spatial position of the spheres so documents whose similarity is higher than a predefined threshold are located closer in space, independently of the previous clustering (see Figure 1.c). Then, users can determine at a glance which documents are similar even if they belong to different clusters. This effect is generated by adding another multiplier to the attraction force between pairwise documents if and only if the distance between them is lower than the mean distance between any other pairwise documents.

2.3. A Gesture-based Interaction Paradigm

As for our daily tasks we use our hands, we think that the manipulation of 3D environments is also more intuitive using them. Therefore, we propose a gestural interaction paradigm where the index and the thumb, equipped with solitary markers, are optically tracked to obtain their spatial position and the distance between them. Both the position and the distance are used to infer which of the three recognizable gestures does the user: maximal distance (thumb and index spread, see Figure 1.d), medium distance (relaxed fingers, see Figure 1.e) and minimal distance (thumb and index pinched, see Figure 1.f). To adapt distances to each user, an initial and directed training phase is carried out. Fingers are virtually represented by two 3D fingerprints (see Figure 2.a), that are immediately identified by the user as representations of his fingers because of the realtime feedback received in terms of position, inclination and rotation while the real fingers move.

In order to test our approach, we have implemented a prototype system with three types of interaction techniques: navigation, selection and control state of spheres. In the first case, we offer three navigation mechanisms allowing the user to reach any point of the environment and to visualize the collection with the desired detail: horizontal and vertical translation of the environment; rotation of the environment with respect to a fixed point; and zoom in and out within the environment. Associated to the zooming mechanisms, and to avoid cognitive overload in the visualization, keywords get visible only when the distance between the spheres they are attached to and the position of the user’s avatar (fingerprints) is under a defined threshold. The second group of techniques deals with the selection of one or more documents, as our system is intended both for individual document search and global exploration of a collection of documents. The first one
permits selecting one of the documents, whereas the latter allows selecting all the documents that belong to the same cluster. For increased flexibility, the system allows making incremental selections, that is adding new documents and/or groups to the current selection. We also provide the possibility of deselecting a group that has been previously selected. At last, the system offers eight different functionalities for manipulating the collection, implemented as pop-up widgets (see Figure 2). To avoid undesired executions of these functionalities during navigation, widgets have to remain activated by the user for at least 2 seconds in order to become operational. Most of the widgets have a binary behaviour and their associated functionality is executed just by activation. The first case is jumping between groups in order to focus the visualization over the centroid of one group. If more than one group has been created, every activation results in focusing on a new group, which is a great help when navigating and exploring over groups and not over individual documents. In addition, the system offers four control functionalities that require a previous selection of a document or a group: consult the content of a document, invert the selection, clear the selected documents or groups from the visualization or recover the last documents or groups cleared. The last three functionalities require a numerical value, so, besides activation, their widgets require the user to indicate a number, that is also provided by finger movements. This is the case of the dynamic adjustment of the number of groups for the clustering algorithm (see Section 2.1), and the adjustment of both force modifiers added in the FDPFC approach (see Section 2.2).

3. Experimental Evaluation

3.1. Experiment Design

A user evaluation was designed in order to evaluate the impact of the modifications introduced by the FDPFC and of the gestural interaction techniques proposed. An experiment was conducted in which one independent variable, the type of 3D mapping method, was manipulated with two possible levels: Force Directed Placement (FDP), and Force Directed Placement with Force Control (FDPFC). A total of 36 subjects were selected for the experiment and randomly assigned to one of the two experimental treatments. All subjects were computer science students or professors, so they were assumed to be

![Figure 2. User interface: a. Virtual fingertips; b. Delete; c. Invert; d. Undo; e. Inspect; f. Document force adjustment; g. Groups adjustment; h. Group force adjustment; i. Jump](image)
able to understand the abstraction applied in the 3D representation of document collections, and to quickly learn and apply new interaction techniques. Ages of the participants ranged from 21 to 40, with 25 males and 9 females. The measured dependent variables were: Precision in the mental model resulting from exploration (M), Effectiveness in search (E), and Satisfaction (S). The null hypotheses were: H01) The mean precision in the mental model obtained from exploration is the same for FDP and FDPFC; H02) The mean effectiveness in search is the same for FDP and FDPFC; H03) The mean satisfaction is the same for FDP and FDPFC. Statistical t-tests were applied to evaluate these hypotheses.

Precision of the mental model was evaluated after an exploration task. The experimental document collection was manually composed by intentionally selecting documents within a set of predefined topics. Precision of the mental model was measured by presenting, after some minutes of free exploration, a brief thematic questionnaire with nine questions in the form "Were there any documents in the collection talking about topic?" A 5 levels response scale was presented with the meaning -2:”definitely no”, -1:”probably no”, 0:”I do not know”, 1:”probably yes”, and 2:”definitely yes”. The reference topics for the nine questions were Ontologies, Buildings, Bridges, Animals, Pedagogy, Software Agents, Memory, Art, and Civilizations. The hypothesis was that the availability of force control mechanisms in FDPFC would allow the subjects to construct a more precise model of the thematic structure of the collection, particularly by strengthening the intra-cluster attraction force.

Effectiveness in search was measured via a search task in which the participant was asked to find within the collection a representative document talking about a particular topic, as well as two additional documents that were the closest to it in their thematic content. The best possible solution for the task was pre-calculated. Taking the inter-document distance matrix, the effectiveness in search was computed as the distance between the solution document provided by the experimental subject and the optimal solution. For the evaluation of the selected neighbors we decided to count the number of best neighbors provided. The hypothesis was that the availability of force control mechanisms in FDPFC would allow the subjects to find a higher quality set of documents, particularly by strengthening the inter-documents attraction force. We hypothesized that the increase in quality would be more significant when considering the neighbors than just by comparing the best representative.

Finally, Satisfaction was evaluated by a questionnaire at the end of the experience. 15 questions were common for both experimental groups, with 3 additional questions for the FDPFC group evaluating the perceived usefulness of the force control capabilities, and 2 additional questions for the FDP group trying to find out the usefulness they anticipated for these features. The questions offered a 5 levels response scale with the interpretation 0:”I strongly disagree”, 1:”I disagree”, 2:”Neutral”, 3:”I agree”, and 4:”I strongly agree”. There were four open questions investigating previous user knowledge of document collection exploration and search tools, and requesting a general evaluation of positive and negative aspects.

In the test arrangement the subjects were standing in front of a projection screen (see Figure 3.a), with two fingertips (see Figure 3.c) tracked with an optical system (see Figure 3.b). The experiment was conducted along four days, with half an hour assigned for each subject. Each subject went through the following stages: 1) Preparation: Before entering the experimental area a document was handed over in which a general description
of the experiment’s goal and procedure was presented; 2) Training: The participant was led to the projection screen and ten minutes were allocated for the participant to familiarize with the visualization display and interaction techniques by using a training collection; 3) Exploration: The test collection was loaded and the participant was instructed to freely explore it trying to get an idea of the documents’ topics during five minutes; 4) Mental model evaluation: The participant was taken to a separate room and five minutes were left for completing the thematic questionnaire; 5) Search: The participant was taken again in front of the projection screen and the search task was completed in five minutes; 6) Questionnaire: Again at a separate room, the participant was requested to complete the satisfaction questionnaire and was thanked by their collaboration. No time limit was imposed at this stage.

3.2. Results and Discussion

When evaluating precision of the mental model, if a topic presented to the user is certainly included in the collection, the correct response is considered to be 2, while it should be -2 if the topic is not in the collection. Any other response is considered incorrect. The topics that were certainly in the collection (Ontologies, Bridges and Memory) were noticed by the majority of the users (with more than 10 correct responses out of 18 participants in each topic and experimental group -FDP and FDPFC-, and a mean of 13 correct responses in both cases). If we consider topics that were not included and were semantically quite far from other topics in the collection (Animals, Art and Civilizations), users found it difficult to assert with certainty that the topic was absent (even if they tended to believe the topic was not in the collection), so the percentage of correct responses is low, but we can notice less uncertainty in the group of users who could control the attraction forces (FDPFC), with higher success rates (mean of 4 versus 2.33 correct responses out of 18 participants). In the case of topics that were not in the collection but were intentionally close to other topics in the collection (Buildings, Pedagogy and Software agents), the success rate is the lowest, as expected, with a slightly highest success rate in the FDPFC group (mean of 1 versus 0 correct responses). In both cases, the reduced amount of time left for exploration (5 minutes) as well as the fact that users only had access to a keyword and the abstract of the documents, can explain the uncertainty that prevented them from asserting the absence of the topics. The test on the equality between the mean number of
correct responses in the FDP and FDPFC conditions did not allow us to reject the null hypothesis ($\sigma=0.32$).

In the evaluation of search effectiveness, two tasks were demanded. The mean distance between the users’ selected document and the optimal one was similar in both groups (0.29 for FDP vs 0.26 for FDPFC in the first task, and 0.09 for FDP vs 0.11 for FDPFC in the second one). No significant difference was found between FDP and FDPFC, with both groups being able to find approximately half of the best neighbors.

Regarding user satisfaction, if we analyze the mean response value for each of the Likert format questions, we can see that all means are higher than 2 (the neutral point in the scale), showing a generally favorable opinion. For the overall questionnaire, the mean response value for the FDP group is 3.12, while it is 3.29 for the FDPFC group, slightly better but still the difference is not significant ($\sigma=0.13$). The most remarkably positive opinions (above 3.5 points) were about the usefulness of keyword visualization, the possibility to jump the viewpoint from one group to another, and the possibility to select a group of documents and remove it. The most valued feature for both groups was keyword labeling. In both experimental conditions, users believed they could make more benefit from the system with more training and experience.

Regarding the features that were available for the FDPFC group but not for the FDP group it is interesting to see that the participants of the latter one did not value much the possibility to include them, despite not having any experience with them (with mean value of 2.31, close to 2-neutral), while the group that effectively had the opportunity to enjoy the features certainly valued much more their usefulness (with a mean of 3.32).

4. Conclusions and Future work

We presented some contributions towards a more intuitive and effective interaction with 3D visualizations of unstructured document collections. A prototype system has been developed fully automating the information visualization pipeline. A modification of the classical FDP algorithm has been proposed (the FDPFC variant) to determine the spatial position for each document’s spherical representation in the 3D space. This modification allows the user to control the attraction forces to be applied among documents so that group separation or inter-document similarity can be visually enhanced, as required for the task at hand. A full set of interaction techniques have been implemented through a new gesture-based interaction paradigm based on tracking the position of two hand fingers (index and thumb) by using solitary markers. Navigation, selection and manipulation of both individual documents and document clusters, as well as adjustment of the visualization settings (number of clusters and attraction forces), can be achieved just with very basic hand movements and a set of reactive zones in space.

After experimentation FDPFC seemed to help decrease the uncertainty about the thematic structure of the collection, although no conclusive evidence was found that FDPFC allows the construction of a more precise mental model. Partly this could be explained by the fact that the exploration task was the first experience of the users with the system after a really very brief training period (just ten minutes), and the participants may have felt overwhelmed by the new way to interact with the system, the number of functionalities offered and the tasks required. In fact we observed very limited use of the force control options during exploration. We believe that a longer training period in
which users have the opportunity to really understand the effect and possibilities of each interaction mechanism could result into more effective use of the force control features and a more significant effect on the precision of the mental model.

Regarding the user satisfaction and subjective perception, keywords were the most valued option. This is probably due to the fact that it was the most evident way to look for the required documents in the search task. Some users applied an almost exhaustive search strategy, inspecting every document’s keyword, or even explicitly consulting all documents’ contents. This strategy could suffice in our limited-size experimental collection, but we believe that if faced with much more populated collections, users would start to realize the usefulness and higher efficiency of alternative manipulation techniques.

In our future work we plan to experiment with bigger and more complex document collections and with a longer and more explicit training period that really helps users to become proficient in the use of the interaction techniques, to face users to situations in which the application of the new interaction techniques really becomes a worthy alternative to more evident strategies, and to avoid results to be biased by the initial user’s disorientation.

References


Cover sheets considered harmful

Emma L. TONKIN a, Stephanie TAYLOR a and Gregory J. L. TOURTE b

a UKOLN, The University of Bath
b The University of Bristol

Abstract. The spread of the cover sheet is a divisive phenomenon. Their appearance is geographically bound and their content situated in the local political and financial context. In this article we discuss the arguments for and against the cover sheet in its guise as a fixture on institutional repository preprints, exploring the issue through statistical information gathered from survey material. We lay out the reasoning behind the use of cover sheets in the United Kingdom and discuss their prevalence and the underlying trends.

Keywords. institutional repositories, cover sheets, survey, copyright, text analysis

Introduction: What are cover sheets?

Cover sheets may appear an unlikely source of controversy. Simply put, they are additional pages added to a resource in a digital repository, almost always prepended to the first page rather than appended as an appendix to the document. The use of cover sheets is often associated with the rise of branding in institutional repositories—that is, they are seen as the solution to the marketing challenge faced by institutional repositories seeking to render the scope and usage patterns of their activities more visible in light of the need to retain funding in a difficult research environment. Institutional repositories, however, face many other challenges, including a [perceived] need to retain a visible copyright statement on each file and to provide user-accessible metadata. In a world in which the majority of resources are accessed directly through links to the resource, rather than through the intermediary of the institutional repository, a cover sheet is the most visible and arguably the most accessible way for institutional repositories to provide useful information to the reader.

Guidance for the institutional repository manager in the United Kingdom has generally been positive on the subject of cover sheets. For example, Brace [3] wrote that the benefits of consistently constructed cover sheets applied either to a subset or a full set of repository objects included the availability of detailed version information (assuming, implicitly, that the cover sheet information set included versioning data) and the possibility of linking to other information such as copyright details. The only downside identified in this instance was the time and resource requirement of the process of cover sheet creation; the solution identified was the use of an automated system for cover sheet creation.

Internationally, viewpoints on cover sheets vary widely. Large archives such as JSTOR provide work behind a cover sheet which provides a standardised view of various information about the document, including citation information. Sites such as Arizona’s
DLIST do not embed a full-page cover sheet, although some citation or versioning information may be provided within some documents. The QUT repository in Australia, by comparison, routinely places a cover sheet at the beginning of PDF content. Establishing the prevalence of this practice worldwide would be a major task if completed manually and is perhaps best achieved through reliable, automated analysis.

It is perhaps worth identifying a few uses of cover sheets:

- Content submission cover sheets: for example, grant applications, thesis submissions, even homework submitted by students may be prefaced with a cover sheet identifying the submitter and some basic information about the document. These are a form of process management aid, and inherit from the tradition of the print-out cover sheet, which was widely used in many institutions to identify documents when printed.
- Presentational cover sheets: some types of document, such as reports, may be wrapped or prefaced with an appropriate cover sheet for the purposes of imposing a standard presentation on institutional outputs.
- Cover sheets as an aid for the researcher: having identified and printed or copied the document to the local machine, the researcher has lost potentially useful contextual information about the document. Placing that information within the body of the document means that it is, in principle, retained.
- Cover sheets as a means of author identification: some academic authors feel reassured that their intellectual property and identity as the author of a piece of work is more securely asserted by the use of a cover sheet. In such cases, the cover sheet can be perceived as providing the same author information as a traditional journal article.
- Cover sheets for data papers [9]: A ‘data paper’ in this context is proposed to be a type of paper that presents ‘a neutral, argumentfree platform from which to draw scientific conclusions’. The purpose of a cover sheet in this context is the provision of essential indexing information alongside ‘a package of narrative and references that captures the entire data product’ (for example, links to the full dataset).
- Cover sheets as a visual reference for copyright and permissions data.
- Cover sheets as a branding exercise.

Cover sheets and copyright

Copyright is a major guiding force behind the use of cover sheets. Matthies and Frisby [10] describe the use of cover sheets as an integral part of the permissions process. A poster presentation by St Andrews [2] regarding the provision of an electronic theses service describes the use of cover sheets not only for branding of the full-text documents, but also to provide relevant context (‘anchoring’ the full text) and to give reassurance to postgraduates that their work is adequately protected against unauthorised reuse or plagiarism. In other words, there is a perception that cover sheets are either required or beneficial as a visible indicator that a copyright triage/assessment process has been undergone and that any readers will be informed of the copyright status of the work.

In the UK, the SHERPA/RoMEO project [6] provides a series of conditions laid down by publishers for self-archiving. In many of these cases conditions for author self-archiving may include the use of a statement of some kind (i.e., published source must
be acknowledged, a DOI must be provided, etc.). Publishers do not typically specify that this information is to be provided in a cover sheet embedded into the document itself.

**What’s wrong with cover sheets?**

With the exception of cases in which specialised templates are used (such as certain formal report delivery processes, in some cases thesis submissions, and so forth), the use of cover sheets in institutional repositories is almost solely associated with files encoded in fixed-layout formats, particularly PDF. This is probably due to the relative ease and reliability by which these formats can be manipulated on the command line using tools such as pdftk and pdftools. This manipulation is often visible, since cover sheets frequently use a different paper format/layout to the item itself (e.g., letter format vs. A4). The formatting discontinuity is visible to the user and to the printer, and is wholly unjustifiable since there is no reason why, in the cases of common paper formats, a corresponding cover sheet format cannot be generated.

The Version Identification Framework [14] lists a number of pros and cons of cover sheets from the perspective of an enthusiast. The pros listed included

- **Uniformity** The object continues to be identified clearly, including its version status, even if it is removed from the repository itself.
- **Detail** All of the types of essential versioning information [can] be used.
- **Linking** To other repository information, such as policies, that might be related.

The downsides identified included the time and resource commitment required to add cover sheets, which they propose should be offset with automated alternatives, perception of interference (introduction of potentially unwelcome branding, relegation of content to second page), problems with page numbering and, finally, preservation issues. These issues are covered in the following statement: ‘Some may view the object as representative of how the research process stands at the time of deposit, and that altering it, even in such a ‘behind the scenes’ way, is a threat to the integrity of the work in an archiving sense.’

The VIF proposed solution to this last issue is the use of the PREMIS data dictionary [12] for preservation metadata to store information about any changes made. However, the proposed solution requires the identification of what constitutes useful information and what should be retained. For text mining purposes in particular, it is easy for minor changes to significantly alter the usefulness of the file for a given approach.

Consider for example a scenario in which a researcher is attempting to make use of PDFs retrieved from repositories in order to perform author name/identity disambiguation upon a document set: that is, the researcher is seeking to make use of available contextual information in order to tell two people called ‘John Smith’ apart. Contextual information typically used for this purpose may include location of publication, classification of the text, subject of the text, references contained within the document and so forth. A researcher could equally choose to look at the structure of the document itself and attempt to tell individuals apart by the platform that they use to generate their documents (OpenOffice versus Word; l\LaTeX on a Mac; printer driver used to print to PDF, Calligra and so forth). It is a simple feature set, easily accessible through manipulation of the PDF format using tools such as pdftk or exiftool, but potentially an important one—but this information is usually removed when a cover sheet is generated. As with
the case of paper size discontinuity, there is no theoretical reason why this information must be lost, since the tools typically used for this purpose are capable of preserving and manipulating metadata.

Consider a second researcher who is looking to index all the documents that he/she has available using an automated indexing tool. The PDF cover sheet presents some difficulties in this context too, since indexing typically picks up the features of the cover sheet rather than the document itself, which may or may not contain the correct information formatted as the indexer expects. Indirectly, the effects of this can be seen on Google Scholar today, since the service makes use of automated indexing as a potential data source (see [5]). The result is that the indexing picks up information from the cover sheet and may, depending on the query, display it in place of more relevant information from the structure of the paper (see fig. 1). In an earlier study, Hewson and Tonkin [7] identified cover sheets as a potential problem for automated indexing and retrieval.

In most cases, Google Scholar is smart enough to resolve this issue satisfactorily. However, the recent spate of articles discussing the low discoverability of articles on institutional repositories [11, 8, 1] raises the question of factors influencing discoverability of articles in repositories; it is perhaps reasonable to wonder whether cover sheets have a positive, negative or neutral impact on the discoverability of items in that repository.

1. Usage of cover sheets: a survey

The rest of this paper is given over to establishing opinions of and usage patterns relating to cover sheets. There are several viable approaches towards this goal. Some are technical, such as the (complete or partial) spidering of institutional repositories for files which can be parsed to identify cover sheets. Initially, we chose to request information directly from repository and CRIS/RIM system managers by means of a questionnaire-based approach. Our primary goals were to elicit the primary motivations for the use of cover sheets, the prevalence of cover sheets in the (predominantly UK) repository landscape, and to identify the purposes for which cover sheets are currently applied. We felt that this would permit us to approach an evaluation of the effectiveness of the approaches taken to fulfil these requirements, allowing us to explore the question of whether and in which contexts cover sheets are effective or, to use Dijkstra’s famous 1968 snowclone, may be considered harmful [4].
Figure 2. Comparing survey findings with other data sources, such as ROAR, RSP and RepUK

1.1. Method

In order to investigate the perspectives of repository managers on the issues and practice surrounding cover sheets, we made use of a quantitative survey method. We designed a short survey with the intention of learning about digital library and resource management system administrator viewpoints on the subject of the usage, benefits and drawbacks of cover sheets. A qualitative discussion (interview) approach was taken initially with a small subset of repository managers, in order to enable the authors to identify the key issues and concepts to cover during the survey itself, which was initially piloted with a further subset of subjects.

The resulting three-page questionnaire is separated into three sections. Questions 1 and 2 (page 1) explore the repository system used and the use, if any, of cover sheets. The central section discusses managers’ views on cover sheets; respondents who answer in the negative to Question 2 complete the survey at this point. Those indicating that they make use of cover sheets instead complete an analogous section discussing the creation and use of cover sheets.

The survey was circulated to three mailing lists with a UK focus, including the JISCMAIL Repositories list, and was announced on Twitter. It was kept open for a total of four days.

1.2. Results

Overall, 88 respondents filled in the survey. In order to establish the reliability of the survey responses we note firstly that the overall size of the population of UK academic institutional repository managers is proportional to the number of academic repositories. These are variously estimated by the repository indexing service OpenDOAR as summing 209, by ROAR as 250, by RSP as 86 (of which the RSP project [13] publishes a full description of 83) and by RepUK as 150. This variation probably relates to variation in the populations targeted by each service, since RSP includes neither
further education colleges nor, geographically, does its coverage include Northern Ireland. The results given here are subject to an approximate ±7 percent margin of error at the 95% confidence level. Comparing the surveyed data for institutional repository platform usage shows that the survey responses are comparable with other data sources’ usage estimates (see Figure 2), with some variation resulting from the survey method, although repeated methods ANOVA shows no statistically significant difference, $F(8,3) = 0.194, \ p = 0.899$.

Anecdotally, it is suggested that Bepress is more widely used in further education in the UK context, whilst ePrints may attract less ‘hacker culture’ than Fedora, the incidence of which is anomalously high in our dataset. Whilst the comparison of these platform usage estimates deserves further discussion in another forum, it is adequate for the purposes of this paper to note that ePrints users in particular are likely to be somewhat underrepresented amongst our respondents.

1.3. Reported usage of cover sheets

57% of respondents responded that their repository included cover sheets on documents. An additional 11% of respondents stated that they made some use of cover sheets, used them for a subset of documents, made use of unconventional cover sheets (for example, placing the coversheet at the back of the document) or were considering making use of cover sheets in future. Thus in aggregate over two thirds of respondents either use, intend to use or instruct repository users to add cover sheets.

1.4. Metadata included within cover sheets

Survey responses to the question ‘what information do cover sheets contain in your repository system?’ presented a spread of pieces of information of which the most popular are document metadata such as title and author, copyright, citation and institution
information, followed by persistent IDs, original venue of publication and the status of the publication (see Figure 3). Cover sheets are overwhelmingly used on categories of full-text documents only, with a few respondents indicating that they were also used on metadata-only records.

1.5. Creation process of coversheets

Both automated and manual processes may coexist in a single repository, depending on the context (i.e., it may sometimes be preferable to generate ‘standard’ sheets for a collection automatically, whilst in other cases a manual process is necessary). 49% of repository managers reported making use of an automated process for coversheet creation, of which 8% report the usage of a batch processing approach based on available metadata. A manual process is used in 53% of cases.

In summary, over half of the repository managers surveyed report that they generate the sheets manually. Cover sheets almost always contain author information, title information and copyright information, and the majority of respondents additionally indicate that citation information is provided, alongside a persistent URL and original place of publication as additional metadata elements provided in cover sheets. In the vast majority of cases, cover sheets are used only on full-text records.

1.6. Motivations for use of cover sheets

Respondents who made use of cover sheets were asked about their primary motivations for doing so (see Figure 4). The most popular reasons were the addition of institutional/departmental branding and document provenance information, followed by citation information and version information. Although copyright concerns were not provided as an option, around a third of this group identified copyright concerns in a write-in field of the survey, several indicating that the provision of a statement set by the publisher is a requirement for allowing self-archiving or that the provision of a cover sheet is a risk-management strategy.

Respondents who indicated that they do not make use of cover sheets were also surveyed about their views. The majority (over 60%) indicated that they considered cover sheets useful for citation, institutional and copyright indication. Relatively few felt that author information (39%), persistent URL (34%), title, original place of publication (30%) or status of publication (22%) were of use. Around a quarter of these respondents indicated that cover sheets were useful for none of these purposes.

When asked for the primary reasons behind the decision not to implement cover sheets, around a third identified a lack of demand from repository users and/or that the question had not arisen. Other issues mentioned included a lack of support from the repository platform, the scalability of the process/the resources involved, both technical and administrative, low priority and, interestingly, identification of alternative strategies—one respondent noted that they redirect incoming ‘deep’ links to the repository item record, requiring the user to click through the item record page, since this exposes all of the information that would be present on the cover sheet.

Text mining or Google indexing concerns were identified by only one survey respondent, and preservation concerns by none.
2. Discussion

Overall, well over half of all respondents indicated that they make use of cover sheets in at least a subset of cases. The results of our survey clearly demonstrate that there is significant ‘buy-in’ among repository managers to the concept of the cover sheet as an instrument for compliance with copyright concerns. Indeed, a number of respondents view them as a necessary condition for compliant archiving of preprints. A puzzling aspect of this is the observation that, even in repositories that make extensive use of cover sheets, items in formats that do not permit easy prepending of cover sheets are nonetheless made available, suggesting that they are not a requirement for distribution in at least a subset of cases.

The issues resulting from this practice can usefully be separated into technical and general issues. Of the former, metadata issues relate to loss of evidence and are thus placed in the domain of preservation concerns. Text mining issues relate both to the loss of metadata and potentially document structural data and to the imposed requirement of completing an additional stage of identifying and removing coversheets before proceeding with the task at hand. Layout and formatting issues result in minor inconvenience to the user when viewing and printing the document and are a consequence of incomplete cover sheet generation (i.e., a good cover sheet generation function should be able to detect the format of the original and respond appropriately).

General issues for the repository manager include an ongoing need for technical effort, time/resource requirements, scalability issues and difficulties encountered when upgrading or changing platforms. Repository contributor/user viewpoints on the cover sheet may, as previously indicated, not be straightforward to characterise, as they are viewed as anything from a waste of paper to a useful protective measure; we will not attempt to cover these in more detail here, as they deserve detailed study in their own right.

For a repository manager looking for alternatives to the use of cover sheets, the following options have been explicitly identified by some respondents and through interviews:
placing relevant information on the repository item page, instead of altering the
item itself
• forcing the visitor to click-through the associated repository item page,
• embedding relevant data within the document metadata itself
• use of a smaller and less intrusive repository ‘stamp’ somewhere on the document,
  enabling branding and provision of relevant information
• provision of unaltered versions of the documents contained within the repository
  for access by means other than direct browsing over the Web, permitting those
  with a use for unaltered versions of the documents to retrieve the originals.

Of these, the third and fourth are as achievable within PDF as the application of
cover sheets themselves (and are no more applicable to other file formats such as DOC
files). The second resolves part of the issue, ensuring that users view the copyright state-
ment/repository branding, but if that document is distributed elsewhere, it will not con-
tain relevant embedded data.

It is perhaps worth noting that whilst it is broadly understood by the majority of
survey respondents that cover sheets often indeed contain useful information, the use of
the cover sheet to protect a document’s copyright attribution is not supported by the tech-
nical application. Cover sheets may easily be removed in bulk or simply by using ‘print
to file’; on the technical level they provide significantly less of a challenge to remove
than a watermark or a stamp. A visible cover sheet provides only the reassurance that
the reader may be assumed to be aware of the information contained within it. That said,
the cover sheet seems to have become an agreed convention for fulfilling the compliance
requirements of all parties with an interest in rights management. As such, it can be said
to have become a ‘quick win’ way of dealing with these issues for busy IR managers.
Establishing and maintaining a good working relationship with authors and rights hold-
ers is an essential part of the job for IR managers, so any quick, easy and agreed way of
dealing with this potentially tricky area is welcomed into their busy work flow.

3. Conclusions

On the evidence presented here, the adoption of cover sheets is widespread in the UK,
although far from uniform. There is a strong argument for the use of some mechanism to
present relevant information, as publishers often request that this is done; however, it is
not clear that this information must be embedded within each digital object. If this proves
to be the case, it will raise questions about the provision of other types of document such
as data files in formats that do not support a cover sheet.

In conclusion, the authors recommend a number of possible steps for repository
managers. Consider alternative ways to brand your documents, such as using a water-
mark or a ‘stamp’ such as that used by ArXiv. Explore the possibility of holding metadata
currently contained in the cover sheet elsewhere in the document, such as in metadata
fields specific to that file format. Review the performance of your cover sheet generation
process, if you use one, and discuss any technical issues identified such as loss of
metadata or formatting issues with the repository vendor or support community. Finally
ensure that the original versions of files are retained, as you may wish to change your
approach in future.
Future work in this area should include a broader survey of cover sheet use worldwide, as well as focusing in some detail on the opinions of repository users and contributors. However, the authors would like to emphasise once again that repository managers don’t make the rules; the repository manager is tasked with identifying and applying an appropriate compromise between the concerns of the different stakeholder groups involved, which is not a trivial undertaking.

References

Me4DCAP V0.1: a Method for the Development of Dublin Core Application Profiles

Mariana CURADO MALTA a, Ana Alice BAPTISTA a

a University of Minho, Algoritmi Center, Portugal

Abstract. Recent studies show that there is no method to develop a Dublin Core Application Profile (DCAP). A DCAP is a very important construct to implement interoperability, therefore it is essential to have a method to be able to develop such a construct, in order to give DCAP developers a common ground of work. This paper presents the first version of a method to develop Dublin Core Application Profiles (Me4DCAP V0.1) that has been developed in a PhD project with a Design Science Research (DSR) approach. Me4DCAP was built having as starting point the Singapore Framework for DCAP and shows the way through the DCAP development. It encompasses a group of pre-defined interconnected activities, explicitly states when they should take place, what techniques could be used to execute them and what artifacts should result from their execution.

Keywords. Semantic Web; Metadata; Dublin Core Application Profile; DCAP; development of DCAP; Me4DCAP; Method; Design Science Research

1. Introduction

The Semantic Web, or Web of Data, has technologies that “enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies”[1] that started to emerge in 1999. The Semantic Web is about common formats for integration and combination of data from different sources [1]. Metadata, both in its use and in its definition and description, is present at various levels. Metadata is data that describes resources with information [2] and that follows well-defined rules of metadata schemes. A metadata scheme is a set of “metadata elements designed for a specific purpose, such as describing a particular type of information resource” [2, p. 4].

The Dublin Core Metadata Initiative2 (DCMI) created new instruments so that those involved in the definition of metadata descriptions could speak a common language. These new instruments appeared with the aim to adapt the metadata community to the transformations the Semantic Web brought. The Dublin Core Abstract Model (DCAM) [3] appears with this purpose: it is a model developed by DCMI, for DCMI syntax spec-
ifications, that presents the components and constructs used in DCMI metadata. One of these constructs is the Dublin Core Application Profile (DCAP), “a generic construct for designing metadata records” [4]. The Singapore Framework for Dublin Core Application Profiles’ recommendation - c.f. [5] - defines the rules to build a DCAP.

A DCAP is a very important construct to implement interoperability, therefore it is essential to have a method\textsuperscript{3} to be able to develop such a construct, in order to give DCAP developers a common ground of work. For the time, being the only guidelines available to develop a DCAP are stated in the Singapore Framework and the DCMI Guidelines for DCAP - c.f. [5] and [4] - and they are too brief. In fact, recent studies show that there is no formal method to develop a DCAP [7]. The absence of guidelines showing life-cycle with standardised activities, as well as a set of well-defined design criteria, with defined techniques, make a DCAP development rather a craft than an engineering activity. Therefore, it is imperative to have a method for the development of DCAP.

The work detailed in this article is framed in a PhD project that has as goal to develop such a method. In the article we present a condensed description of the first version (V0.1) of the method for the development of DCAP (Me4DCAP). A report with a detailed description of Me4DCAP is available at http://hdl.handle.net/1822/23537. This research is still in progress. This article is organised in 4 sections. In Section 2 we present the Design Science Research methodological approach. In Section 3 we present the first version of our results, i.e., Me4DCAP V0.1. Finally, conclusions and future work are drawn in section 4.

2. Design Science Research methodological approach

Our work is based on a design science research (DSR) approach. Design Science aims at the development of innovative artifacts that solve real-world problems [8], thus "Design Science is inherently a problem solving process“ [9, p. 82]. An artifact is something that is artificial, constructed by humans [10], a project using a DSR methodology produces artifacts that can be either constructs, models, methods, or instantiations of one of these 3 artifacts [11]. "Artifacts must be improved upon existing solutions to a problem or perhaps provide a first solution to an important problem“ [10, p.6]. Our research project will develop a first solution to a problem: a method for the development of DCAP (Me4DCAP) to provide the metadata community with a method that, to the best of our knowledge, does not exist.

According to A. Hevner [11] DSR has 3 cycles: (i) the "Relevance Cycle“ that works in the "Environment"; (ii) the "Design Cycle“ that works in the core activities of building the artifact, and; (iii) the "Rigor Cycle“ that works in the "Knowledge Base" of scientific theories.

In the Relevance Cycle the “Environment“ supplies the research project with the needed requisites and the application context, and "defines acceptance criteria for the ultimate evaluation for the research results“ [11, p. 89].

In the Design Cycle DSR "calls for multiple iterations (...) before contributions are output into the Relevance Cycle and the Rigor Cycle“ [11, p. 91]. These iterations are

\textsuperscript{3} A method is a selection of techniques, the control of their usage and the integration of the obtained partial results [6].
cycles of construction and evaluation, and "these evaluation moments are performed in laboratory and experimental situations" [11, p. 91].

In the Rigor Cycle DSR uses as input for the Design Cycle the knowledge base "of scientific theories and engineering methods that provides the foundation" [11, p. 89] and the DSR project feeds back the knowledge base with new artifacts and "the experiences and expertise that define the state of the art in the application domain" of the DSR research project.

3. Description of Me4DCAP V0.1

3.1. The DCAP development work team

Me4DCAP defines 4 types of stakeholders that interact in the DCAP development process: Managers, System Analysts, Metadata Programmers and Final Users. By Manager Me4DCAP means a manager of an organisation that has a Web system that has implemented or will be implementing the DCAP in development. By System Analyst Me4DCAP means a specialist that has technical skills in data modeling and in requirements elicitation, this person should also have some skills of group management. A Metadata Programmer is a specialist in metadata that can use the DSP [12] and RDF languages, and understands the Semantic Web concepts. By Final User Me4DCAP means a user that works with the Web system that has implemented or will be implementing the DCAP in development. It should be noted that the multidisciplinary team is very important and should be respected for the success of the task of developing a DCAP.

Since the work-team is composed by persons from different backgrounds having different skills, it is very important to build a Glossary. This should be done from the beginning of the DCAP development process. A Glossary is a text document with the keywords (and its description) used in the DCAP. A Glossary is used to define important words commonly used by the work team while constructing the DCAP. In multidisciplinary teams it is important that the members of the work team speak all a common language, as it avoids misunderstandings and improves communication [13].

3.2. Me4DCAP approach

Me4DCAP was built having as starting point the Singapore Framework for Dublin Core Application Profiles (c.f. Nilsson et al. (2008)). According to the Singapore Framework, a DCAP is composed by:

- Functional Requirements (Component Stage 1);
- Domain Model (Component Stage 2);
- Description Set Profile (Component Stage 3);
- Usage guidelines (optional) (Component Stage 4);
- Syntax guidelines (optional) (Component Stage 5).

The starting point in the Knowledge Base is the Rational Unified Process (RUP) (c.f. Kruchten (2004)). Me4DCAP establishes the way through the DCAP development: it establishes when activities must take place, how they interconnect, and which artifacts
they will bring about; it also suggests which techniques could be used to build these artifacts.

The development of a DCAP is an iterative process by stages, each stage being built on the results of the previous stage. This set of stages is our starting point for the definition of Me4DCAP. Me4DCAP defines a way for the construction of each component of each Singapore Framework stage. The components of each stage are called Component Stage and each one is identified by a number.

Me4DCAP has 4 phases (see figure 1): Scope Definition, Construction, Development and Validation. These phases are traversed along the project development as the Singapore Components Stage are being developed.

![Figure 1. The Me4DCAP V0.1 phases](image)

In the phase “Scope Definition” work planning initiates, its goal is to define DCAP application scope and to organise the work team. In this phase it is also where it is developed part of the Functional Requirements Component Stage 1. However, the development of the latter is not tight to this phase and overflows to the next one, Construction. In this phase, the Domain Model Component Stage 2 is developed; however, the development of this Component Stage is not, as happened before, tight to this phase and overflows to the next phase, the Development phase. In the Development phase the DSP Component Stage 3 is built. It the climax of all construction done until this moment, since the DSP Component Stage 3 development work is based on the Domain Model Component Stage 2 previously constructed and it is the Component Stage that defines the DCAP in its entirety. Finally in the Validation phase, the developed DCAP is validated. The 2 Guidelines Component Stage 4 and Component Stage 5 are developed throughout the Construction, Development and Validation phases.

3.3. The Me4DCAP life-cycle development model

Throughout a DCAP development, artifacts are being produced to help to reach the aforementioned Component Stages. Figure 2 shows Me4DCAP life-cycle development model mentioning the artifacts that have to be produced and when they should be produced. The Me4DCAP life-cycle development model is iterative. As it is not possible to define all the requirements at the beginning of the DCAP development process, during its development one may feel the need to go back to the previous stage to add missing requirements.
These iterations are at the end of Block 2 to Block 1; at the end of Block 4 to Block 2, or Block 1; at the end of Block 6 to Block 1, and at last at the end of Block 7 to Block 1, which presupposes a new iteration of the whole process. The number of iterations of the whole process depends on the dimension and complexity of the DCAP to be developed. Iterations will end when there is nothing new to “discover” or to add, as for requisites, depending on the results of the validation in laboratory and in production (see sections 3.6 and 3.8).

In order to develop a DCAP, the DCAP developers will need to follow the life-cycle of the development process, building the artifacts that will be used for the construction of the Component Stages (see figure 2). Some artifacts can be developed at the same time; that is the reason for them to be together in the same block. But some can’t be built before others. Figure 3 shows the dependence among the artifacts.

Next section will describe each Me4DCAP artifact and the techniques that should be used to develop them. This description follows the Singapore Framework Component Stages order; they are the center of all development.

3.4. Functional Requirements (Component Stage 1)

To build the Functional Requirements Component Stage Me4DCAP defines the need to develop a set of 4 mandatory artifacts: the Vision Statement, the Work Plan, the Use Cases High Level, and the Use-Cases Model. The first artifacts to be build are the Vision Statement, the Work Plan and the Use Cases High Level. After that, follows the Use-Case Model. The Vision Statement is a document that shows what developers want to reach with the DCAP development. It defines the scope of the DCAP; it is a simple plain
text document with no more than 200 words, describing the boundaries for the DCAP usage. The technique used to develop the Vision Statement should be the brainstorming technique, where all the members of the team should feel free to write ideas in a board (physical board or web tool), followed by discussion. In the end, the set of ideas chosen should be organized in simple sentences.

The Work Plan has as goal the time planning of the project activities; it is the follow up project timing and serves as a guide for the work team of the DCAP development project. The Work Plan refers the timings of each phase as the respective beginning and ending dates, and still the dates when each Component Stage should be ready. It will also be possible for the work plan to include information on the responsibilities of each element of the work team in the phase or artifact in question. The work plan is a text document, a Gantt Chart or any other type of graph or scheme that the work team finds more convenient. The Work Plan should be built by all the members of the work team, and negotiated among them in order to fit all the members of the team time constraints. It is acceptable that the Work Plan has to be modified as the project evolves.

The Use-Cases Model is build after these 3 artifacts previously described are developed. Having Me4DCAP an iterative life-cycle development model, the previous artifacts might have to be revisited more than once, there will be moments the work-team will have to decide to release a draft version of the artifacts to follow the process, being aware that they will be working on those draft versions sometime later in the process.

Use Cases “offer a systematic and intuitive means of capturing functional requirements” [13, p. 37]. The Use-Cases will be used to develop the Functional Requirements
and to understand the objects (and attributes) of the system to be studied. The Use-Cases Model is composed of:

- the UML Use-Case diagram with the actors that interact in the Use-Cases, describing the functionality of the system [14];
- the set of all detailed Use-Cases.

For information on how to build an UML Use-Case diagram see [15] or [13].

“Each use-case must include details about what has to be done to achieve its functionality” [16, p.21] Every Use-Case should be then documented in detail. This documentation should set the sequence of actions - a specific sequence of events that happen in the system - that a system performs to bring added value to a specific actor. An actor is somebody or something (automata) outside the system that interacts with it [14]. An Use-Case description is a flow of events description, and it should be developed using the template proposed by [16]. Every manager member of the work-group will know precisely which are the needs of the system in order to achieve certain objectives of functionality; they should be the persons to identify what are the actions that will bring value to the system. So, the Use-Case description should be developed by the jmanagers, giving them the template of the flow-of-events defined by [16] and a definition of Use-Case (defined in the Glossary). The set of Use-Case descriptions should be written on the board (physical board or web tool), and the work-team as a whole should revise them, with the System Analysts members of the work-team helping managers to clarify ideas.

After having the previously described 4 artifacts developed, the Functional Requirements can be built. This Singapore Component Stage 1 is mandatory. Functional requirements “guide the development of the application profile by providing goals and boundaries and are an essential component of a successful application profile development process. This development is often a broad community task and may involve managers of services, experts in the materials being used, application developers, and potential end-users of the services” [4]. The Functional Requirements Component Stage 1 is a text document, where general goals are mentioned as well as specific tasks [4]. To develop the Functional Requirements the work-team should read, in group, the Use-Cases detailed description to identify which are the functional requirements that the use-cases explicit. Short sentences should be used, and should be written on the board (physical board or on the working web tool). After that, the work-team should identify if there are no repeated ideas of functional requirements on the board. Certain ideas speak more to some work-team members than to others so, each functional-requirement-idea should be distributed accordingly, in order to satisfy every member’s needs specific requirements. Every member of the work-group should write some sentences describing more deeply the requirement-case that is responsible for. In the end of the process, all the requirements-cases should be put together on the board (physical board or web tool), and the whole group should discuss and review the final result.

3.5. Domain Model (Component Stage 2)

The Domain Model is the mandatory Singapore Component Stage 2. It “captures the most important types of objects in the context of the system,” [13, p. 119]. According to [4] “a domain model is a description of what things your metadata will describe, and the relationships between those things. The domain model is the basic blueprint
for the construction of the application profile”. The domain model is build based on the Functional Requirements Component Stage 1 and on the Use-Cases Model artifact described in Section 3.4. The domain model development can also use the help of other techniques depending on the DCAP development context. In situations where access to documentation or the information systems databases is available, it is possible to resort to the Document Analysis technique to define it as well.

Me4DCAP suggests that the Domain Model should be developed using an UML class diagram with details suppressed. The diagram identifies the classes of objects and the relationships among them but the classes’ methods and attributes are omitted since the methods’ definition is not in the frame of a DCAP development, and the attributes will be defined in the ORM Diagram data model (see next section). The Entity-Relationship diagram [17] showing the entities and the relationships among them can be an alternative to the UM class diagram technique (attributes should be omitted).

3.6. Description Set Profile (Component Stage 3)

To develop the mandatory Singapore Component Stage 3 Description Set Profile (DSP) Me4DCAP defines the need to develop a set of 2 mandatory dossiers:

- The Integration Dossier;
- The Validation Dossier (in laboratory).

The Integration dossier comprises 3 artifacts: an Object Role Modeling (ORM/NIAM) [18] diagram data model, a State Of The Art and a Document of Integration. All these artifacts are mandatory.

The ORM diagram is a data model with:

- the classes of objects (defined in the Domain Model);
- the attributes of the classes of objects;
- the attributes’ constraints, such as their repeatability, domain and multi-language option.

In this part of the DCAP development process, every class of object and attribute should have been already described in plain text, in the Usage Guidelines Component Stage 4 (see section 3.7 for details about this artifact). This work should be done by the stakeholders.

After defining the ORM Diagram data model, Me4DCAP defines as next step the application of a metadata scheme property to every attribute of the objects of the Domain Model. The attributes are described, each and everyone, by the existing properties of the metadata schemes of the metadata community. In case of not being able to describe some of the attributes with the existing metadata schemes, those attributes should be described with new properties; these new properties have to be created. According to Baker & Coyle (2009) [4], this process is done in 2 steps:

- To perform a State of the art to existing metadata schemes - that are described in RDF - to find out from the existing schemes which ones can describe the identified attributes. This work should be done by the metadata programmers of the work-team;
- To create new properties - in case there are no properties on the metadata schemes of the state of the art to describe some of the identified attributes. This work should be done by the metadata programmers of the work-team.
The existing information on the State Of The Art and on the ORM Diagram will be used to build a Document of Integration. This Document shows, in a matrix, per line, every attribute and its constraints, described by the properties of the metadata schemes and encoding schemes chosen. This work should be done by the metadata programmers of the work-team. An example based on the DCMI Guidelines (c.f. [4]) can be downloaded from the repository of University of Minho accessible through the URL http://hdl.handle.net/1822/. Me4DCAP defines as next step the execution of the validation of the work done until the present moment of the development process. In order to do that a validation in laboratory is executed; Me4DCAP calls it the “Validation Dossier”. The Validation Dossier comprises 3 mandatory artifacts: a Validation Report, a filled-in Matrix and a filled-in Questionnaire.

A laboratory validation should take place, so as to check:

1. its adequacy to what has been defined in the “Vision Statement” artifact: a meeting of the work-team should take place to evaluate the answer to the defined vision (see Vision Statement in section 3.4). The work-team should make a report (text document) with the conclusions of the meeting and recommendations.

2. DCAP adaptation to the resources that are going to be described by the DCAP: the validation is done through the application of the DCAP to a resource sample. This validation work is done in 2 stages:

(a) Application of the DCAP to a resource sample. The work-team should identify a set of resources that constitutes a trustworthy sample of the application domain of the developing DCAP, and from there, final users, chosen by the stakeholders members of the work-team, and the metadata programmers as resource persons, should complete the validation matrix with data referring to each resource. The matrix template should be simple to fill in, where each element of the metadata is populated with the data that corresponds to the resource. This matrix should be accompanied by the 2 Guidelines Component Stage 4 and Component Stage 5. An example of a validation matrix can be downloaded from the repository of University of Minho accessible through the URL http://hdl.handle.net/1822/. The example used is the same as in the Document of integration (see previous point);

(b) Answer to a set of questions. The final users chosen to do the application of the DCAP to a resource sample (see previous point) together with the metadata programmers of the work-team, should answer to a set of questions to assess the difficulties of the validation process. The goal is to assess if there is data for which the DCAP has no description, or if there are DCAP elements, defined as compulsory, that could not be fulfilled with the information existing in a given resource, or any other type of difficulty or ambiguity. The questions to be asked to the DCAP validators could be like:

i. Could you describe all your data with the available elements? If not, please refer the difficulties;

ii. Were there any DCAP metadata elements left that you could not fulfill? Which? Did this happen for lack of data or because you did not know how to do it?;

iii. Did you have any difficulty in particular to describe your data? Were there any ambiguities?;
iv. Is there anything else you want to add?

According to the results of the questionnaire, the process iterates or follows to the DSP development (see figure 2).

The Singapore Component Stage 4 DSP is mandatory. The task is to detail the metadata developing their design in the DSP language defined by Nilsson (2008). Further information, including implementation examples, can be found in Baker & Coyle (2009) [4]. This task should be performed by the metadata programmers of the work-team.

3.7. Guidelines (Component Stage 4 and Component Stage 5)

These guidelines are not mandatory in the Singapore Framework. Me4DCAP does not make them mandatory but recommends that they are developed since it helps the final users of the DCAP application to apply correctly the properties and constraints.

Me4DCAP defines that the developing of the Usage Guidelines starts at the same time as the Domain Model. DCMI Guidelines [4] explain: “Description Set Profile defines the “what” of the application profile: usage guidelines provide the “how” and “why”. Usage guidelines offer instructions to those who will create the metadata records. Ideally, they explain each property and anticipate the decisions that must be made in the course of creating a metadata record” [4]. For detailed information see [4] and [5].

The development of the Syntax Guidelines needs that the Integration Dossier is developed in a certain stage. This artifact describes “any application profile-specific syntaxes and/or syntax guidelines, if any” [4]. For detailed information see [4] and [5].

The Usage Guidelines can be developed by both types of members of the work-team, the stakeholders and the metadata programmers, since the description of the attributes and classes of objects is information that will have to be filled in by the domain experts. The Syntax Guidelines have to be developed by the metadata programmers since it is a very technical document.

3.8. Finishing the DCAP development

A validation in production of the DCAP should be performed. This process of validation can be done using a log registration technique or observing final-users working with the system that has implemented the DCAP developed. The results of this validation in production should be reported to the work-team in order to review and access the DCAP definitions. If there is new information to introduce in the process, the whole DCAP development process should start from Block 1 (see figure 2), and every artifact should be checked against this new information.

4. Conclusions and future work

A Dublin Core Application Profile is a very important construct to implement interoperability in a community of practice. Previous studies [7] has shown that there is no method to develop such a construct. The absence of guidelines showing life-cycle with standardised activities, as well as a set of well-defined design criteria, with defined techniques, make a DCAP development rather a craft than an engineering activity. Thus, this
is a problem that the metadata community has faced; this community needs a common ground of work concerning the development of DCAP.

This paper presents version 0.1 of a method for the development of DCAP (Me4DCAP), based on PhD research in progress project that uses a Design Science Research methodological approach. Me4DCAP establishes the way through the DCAP development: it establishes when activities must take place, how they interconnect, and which artifacts they will bring about; it also suggests which techniques could be used to build these artifacts. Me4DCAP defines a way for the construction of each Singapore Framework [5] component.

As future work we envision the validation process of Me4DCAP under the scope of the DSR approach; this validation is a very important tool to access the adequacy of Me4DCAP to the needs of the metadata community application domain. These validation process will be done using the Focus Group approach and we expect to find projects in developing DCAP using Me4DCAP V0.1 for feedback and fine-tuning A new version of Me4DCAP (V0.2) will be developed, integrating the information reported from the validation process.

Acknowledgements

This work is sponsored by FEDER funds through the Competitivity Factors Operational Programme - COMPETE and by National funds through FCT - Foundation for Science and Technology within the scope of the project: FCOMP-01-0124-FFEDER-022674.

We would also like to thank Ida Cruz for the drawings of this paper. Last but not the least; we would like to thank Maria Emília for the English review of the paper.

References

COMFy - A Conference Management Framework

Christian CALDERA a, René BERNDT b,1
and Dieter W. FELLNER a,c

a Institute of Computer Graphics and Knowledge Visualization (CGV), TU Graz, Austria
b Fraunhofer Austria Research GmbH
c GRIS, TU Darmstadt & Fraunhofer IGD, Darmstadt, Germany

Abstract. Organizing the peer review process for a scientific conference can be a cumbersome task. Electronic conference management systems support chairs and reviewers in managing the huge amount of submissions. These systems implement the complete workflow of a scientific conference. We present a new approach to such systems. By providing an open API framework instead of a closed system it enables external programs to harvest and to utilize open information sources available on the internet today.

Keywords. conference management system, linked open data, open Bibliographic Data

Introduction

In scientific publishing today most conferences and journals use an electronic conference managing system in order to organize their peer reviewing. Peer review is described as a process of self-regulation by a profession or a process of evaluation involving qualified individuals within the relevant field. Peer review methods are employed to maintain standards, improve performance and provide credibility. In academia peer review is often used to determine an academic paper’s suitability for publication[1].

A typical peer review process for a conference or a journal works as follows: Authors submit their papers in a conference management system. After the submission deadline the reviewers are assigned to these papers. After the reviewing deadline the review results are communicated back to the authors. For all accepted papers the camera-ready version needs to be submitted from the authors. The reviewers are selected from a pool of known experts in the domain of the conference/journal. In case of a conference these experts are usually listed as the international program committee (IPC). For this reviewing process the two most applied types are single-blind and double-blind peer review. Single-blind means that the authors dont know who is reviewing their paper. The reviewer on the other hand knows the identity of the author. In the double-blind review process the reviewer does not know the identity of the authors.

1Corresponding Author: René Berndt, Fraunhofer Austria Research GmbH, Inffeldgasse 16c, 8010 Graz, Austria; E-mail: rene.berndt@vc.fraunhofer.at.
Although there are many criticisms about peer review\cite{2}, the following quote from Mayur Amin from Elsevier\cite{2} describes the current situation:

"Peer review is not perfect, but its the best we have."

One particular point of criticism is the poor referee selection\cite{3}. Especially conferences with a large number of submitted papers experience an enormous time pressure for finding suitable reviewers.

So how can a conference managing system support the conference chair during the reviewer assignment phase? The idea is to utilize information available from different sources about the particular persons in order to find a suitable reviewer or identify conflicts of interest. Especially bibliographic data can be a valuable source of information for finding suitable reviewers. Figure 1(a) shows the image section of data sources related to publications within the Linking Open Data cloud. An example of such a data source is the DBLP\cite{4}, which provides bibliographic information on computer science.

Another bibliographic service is Microsoft Academic Search\cite{5}. Figure 1(b) shows the visualization of a co-author graph from Microsoft Academic Search. A potential reviewer does most likely have an conflict of interest if he is a direct neighbor in the co-author graph of one of the authors.

In this paper we present the current development of the next version of the conference managing system used by the European Association for Computer Graphics (EG\cite{3}). In order to be able to easily extend the system, it is based on exposing an API for managing conferences. This should encourage users to extend the system by developing their own tools and modules.

---

\cite{2} Peer Review at the APE (Academic Publishing in Europe) Conference, Berlin, January 2011
\cite{3} www.eg.org
1. Conference Example: Eurographics Annual Conference

In order to get an insight of the work of a conference chair we take a detailed look at the Eurographics Annual Conference, which is organized by the EG. Figure 2 shows the number of submitted/accepted papers over the last twelve years. The requirement of at least 4 reviews for each paper leads to approximately more than 1000 review assignments. Assuming that the average workload of a reviewer should not exceed 5 reviews means that at least 200 suitable (and willing) persons have to be found. Since the year 2000 the EG uses the MCP system (Managing Conference Proceedings) [6] and the successor SRM (Submission and Review Management) as their conference management system. These system have been especially tailored to support the needs of the EG. In 2012 the system has been used by a total of 25 conferences.

How does the review process work in detail? The submitted papers are distributed to the members of the IPC. Each paper is assigned one "primary" and one "secondary" reviewer. These act as editors for that particular paper meaning they are responsible for finding at least three additional reviewers. Distributing the available submissions to the IPC members has turned out to be the most time-consuming task for the chairs in the last years. In order to support the distribution process the so-called "Bidding-Phase" was introduced with SRM. IPCs are presented a list of all submitted papers (title, abstract, keywords). For each of these papers the IPC member could specify one of the following: "want review", "could review", "not competent", "conflict of interest". Based on these entries the system creates an automatic suggestion how to distribute the IPC members as primary/secondary reviewers. Additionally the IPC members could specify the degree of expertise to the available categories for the papers ("expert", "passing", "not competent"). These values were matched with the author-selected categories for each paper. The weighted sum of both values would then indicate the appropriateness of an IPC member for that specific paper.

Although the process of peer reviewing is unquestioned within Eurographics, over the years valuable input from the chairs in order to improve the process have been made. One of the most discussed issues was the selection of suitable reviewers. Although this weighted sum works well for the distribution, the bidding values have to be entered by each IPC manually. Going through a list of more then 200 titles and abstracts is cumbersome.

Therefore the next version of SRM should use a new approach to use information available by Linked Open Data, especially bibliographic information. With the new sys-
tem it should be easy to interact with 3rd party applications or data sources. It should be easy to harvest the data to create statistics and further usage of this data. Additionally this solution should handle in a similar fashion like the current system.

2. Related Work

This section will give a small overview of some of other reviewed conference management systems:

- **Easy Chair** is a free service for managing a conference. It provides the basic features like uploading a submission and a reviewing system. Further it has multiple conference models a chair can choose from to customize its conference. Besides the models it is not possible to further modify the conference[7]. The review assignment process in Easychair works manual or automatic. When using the automatic mode the Program Committee defines the conflicts of interests and then they specifies which papers they are interested to review. After this is done Easychair tries to create a good matching between the Committee and the papers. [8]

- **COMS Conference Management System** has a one time set up fee for creating a website to satisfy the needs of the chair. This website will be the frontend for their conference management system. Once the homepage is created the chair may define 9 different review fields. The review assignment works is again either manual or automatic. The automatic mode takes the reviewers biddings like in Easychair and creates a matching between the reviewer and the submission. [9]

- **OpenConf** is a php based conference management tool which has again the standard functionality for managing a conference. OpenConf provides the basic conference management tools. They further offer additional modules to add functionality to the program. One of these modules called the bidding module adds the functionality for Program Committee members to define which papers they want to review. After this bidding OpenConf provides some different algorithms to create a matching between the reviewer and the papers. [10]

- **Confious** The Confious system has also the standard features for managing a conference. Confious has like the other systems a automated and a manual reviewer assignment system. But unlike the other systems Confious takes the paper topics into consideration. Authors define which topics their paper is in and the Committee members set their experience in these topics. Then it tries to create a good matching. Further Confious also tries to generate automated conflicts based on the Email and the institute of the IPC and the author. [11]

- **Conftool** is a tool which provides many different languages to manage a conference. Like Confious its automated review assignment takes the IPCs bidding and the paper topics into consideration when creating a review assignment. It also tries to create conflicts like Confious according to the Email, the organization and the surname of the reviewer. [12]

One major challenge in the Eurographics conference is the assignment of reviewers. There are currently two different approaches to accomplish this task. The automated and the manual assignment. With the manual approach the chairs assign reviewers to
submissions from an user-pool. In the automatic approach the program tries to create a
distribution between the reviewers and the submissions.

Current state of the art management systems use some form of bidding system where
the IPC members can specify which papers they can review and for which they are not
knowledgeable enough. Further they have to specify their conflict of interest. They also
might declare their knowledge on the topics of the conference in order to create a better
matching like Confious or Conftool does. Out of this bidding and declaration the program
can compute a good matching between users and submissions.

This system however does not scale very well. Imagine an Eurographics conference
with over 300 submissions (figure 2) where every submission needs at least 5 reviewers.
This would take 1500 users to read 300 abstracts. Most of the time it is not even possible
to obtain such a huge Program Committee. To tackle this problem EG currently divides
this assignment into two steps. At first the IPC members read the abstracts and will then
be assigned to a submission. Their task is then, to find further reviewers which are not
part of the IPC. Since these “external” reviewers are not doing a bidding no information
about their expertise or conflicts are known to the system.

Using open data sources it is possible to support the IPC members in finding the ex-
ternal reviewers, e.g. by identifying conflicts early. Email-based conflict analysis like in
Confious or Conftool cannot detect conflicts between organisations in cross-institutional
projects because of different subdomains. Co-authorship information, e.g. from the Dig-
tal Bibliography and Library Project (DBLP) provides a much stronger indication for
possible conflicts. With the DBLP provided information it might even be possible to cre-
ate a matching between the IPC members and the submissions by using their publication
history making the bidding process superfluous.

3. Framework Architecture

Based on the conference example the concept how the new system works is now pre-
sented. The new system provides an API for managing a conference. The new SRMv2
system communicates with the core layer through this well defined API. The core layer
(COMFy) provides the application logic, while the SRMv2 purely consists of the user
interface. Additionally external programs are also able to communicate with that API.
COMFy itself maintains the conference data and uses the repository pattern in order to
separate the business logic from the database layer. The states of a paper are represented
by a state-machine. The core layer can be extended by additional modules if needed.

The new system is divided into 5 different layers. To lowest layer is the database and
each of the upper layers uses the functionality of the lower layers and adds additional
features to the system. (see figure 3)

**Database** The bottom layer is a relational database. In our case Microsoft SQL
Server[13] is used. This database was chosen for its filestream feature. Normally
files are stored in either the database or in a directory[14]. When stored in the
database the performance to access the files are decreased drastically otherwise
in the directory the transactional consistency is lost. The filestream feature of the
Microsoft SQL Server combines these features by managing the files in database
management system but stores it in a directory which is managed by the database
itself. This way the transactional consistency is guaranteed and the file can be accessed fast via the directory.

Repositories On the second layer there are repositories. The task of the repositories is to provide the upper layers an easy way to access the data in the database. When COMFy queries one of the repositories, this repository maps the request to an SQL statement. When the query is executed it returns the data to the upper layer. It also works in the other direction, so the upper layer can insert new data or update existing entries.

State-machine The second part of the second layer is the state machine. This is the core of the framework. It manages the phases of every submission. When the submission changes its phase it also changes the access rights of different users. For example a author may not submit a new paper once the reviewing process starts. A further integral part of the design with the state-machine is that it should be easily extensible. For example another phase like a rebuttal phase where authors may object to the decision of the chair should be easy to add to the system. The current state-machine can be seen in figure 4.

COMFy This layer contains the business logic of the conference managing system. It exposes these functionalities through a well defined API. It queries the repositories, parses the data and creates the response. It is also responsible for applying the different user roles, e.g. an author does not have access to the reviewer names, etc. It is designed as a model view controller pattern. So the controller takes care of the request, queries the repository, fills the model with the data from the repository and returns it to the view. Depending on the client the requested data can be delivered as XML, JSON or HTML. The next section will give an short overview of the available API functions.

4. COMFy API

The API on COMFy is based on the representational state transfer (REST) paradigm[15], which utilizes the well-known, well-defined methods of the HTTP protocol (GET, POST, PUT, DELETE). This paradigm is based on clean, human readable, hierarchical URLs for accessing the resources. COMFy uses clean, structured URLs to access the requested data. The API calls can be divided in 4 different categories: UserHome, Account, Conference, Submission.
The "UserHome" API calls are used for retrieving information of conferences and submissions which are tied to the user. This way the user can quickly access his own submissions or conferences. The "Account" API calls are used for managing user accounts, e.g. logging into the system, registering or changing the profile information.

The "Conference" API calls are for managing and viewing the conference. These calls are primarily by the chair when setting up the conference.

The "Submission" calls are used for managing and viewing a particular submission. They need the conference identifier because the paper identifier is only unique within one conference. This way it is easy to identify in the URL the conference a submission is in. The calls are mainly used by authors and reviewers. Some of the submission calls can be seen in table 1.

<table>
<thead>
<tr>
<th>API Call</th>
<th>Call Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference/EG2012/Submission/paper1000/Show</td>
<td>GET</td>
<td>information of submission Paper1000</td>
</tr>
<tr>
<td>Conference/EG2012/Submission/paper1000/editSubmission</td>
<td>POST</td>
<td>saves the new information</td>
</tr>
<tr>
<td>Conference/EG2012/Submission/paper1000/assignReviewer</td>
<td>POST</td>
<td>assigns a reviewer</td>
</tr>
<tr>
<td>Conference/EG2012/Submission/paper1000/removeReviewer</td>
<td>POST</td>
<td>deletes reviewer</td>
</tr>
<tr>
<td>Conference/EG2012/Submission/paper1000/reviewerDiscussion</td>
<td>GET</td>
<td>access the discussion forum</td>
</tr>
</tbody>
</table>

Table 1. Submission API calls

COMFy supports currently three response formats: HTML, XML and JSON. By default the sample application SRMv2 (see section 5) generates the HTML view. By adding a data flag to the request the data will be parsed according to the requested format. Such an example request which returns XML looks like: http://localhost/COMFy/Conference/EG2012?data=xml. Examples of the a XML and JSON output are shown in listing 1 and listing 2.
5. Sample Application: SRMv2

The COMFy API encapsulates the core elements of a conference system. But it can’t provide a clear use-case model what steps a user needs to do for a certain task. Therefore SRMv2 is implemented on top of COMFy as one sample application. To illustrate the difference between the API and a particular use-case we take a look at the review assignment use-case. From the API perspective assigning a reviewer is just a call like:

```java
AssignReviewer(ConferenceId, PaperId, ReviewerId, ReviewerRole)
```

The use case for the review assignment can be seen in figure 5(a). This displays the sequence how a reviewer is assigned to a paper from a user point of view. At first the paper has to be chosen. Then the chair has to select the user and set his reviewing role. Currently there are three different of these roles: primary, secondary and tertiaries. After this task it is possible to modify the standard email to create a more personalized email. At last the chair has to confirm the assignment, so the email will be sent and the person gets his assignment which he can accept or decline.

New in this version is that the chair can now access information from DBLP, e.g. it is flagged if the person might have an conflict of interest because of a co-author relationship. In the current prototype the bibliographic data from DBLP is used to help identifying these conflicts. The DBLP provides an API where users can query for authors. Every author in their system has an unique author identifier. After querying an author...
for the author identifier it can be further used to get the co-authors of that particular person. They also provide information about the amount of publications the two authors wrote together. It is also possible with the DBLP API to receive bibtex files of papers. These papers can also be found with the mentioned author pointer. Figure 6(a) shows the information within SRMv2 collected from the DBLP. Access to other sources like Mendeley[16] or Microsoft Academic Search are already under development.

6. Using Linked Open Data

COMFy currently has two different ways to assign a reviewer to paper. The automatic and the manual assignment. The manual assignment can be seen in figure 6(b). After choosing the submission the person who will be assigned to review the paper, COMFy checks Linked Open Data sources like the DBLP if there are conflicts of interest between the reviewer and the authors. A strong link is found, when the full name appears in the coauthor list of for example DBLP. A weak link is found, when the domain name of the Email, the organization or the surname of the authors of DBLP matches.

For the automatic assignment it is necessary for IPC members to complete three steps. At first they are presented with a list of all authors where they can set their conflicts of interest with them. Then they set in which area they are experts in. In their last step they are presented with every paper. There they set which paper they would like to review and which they are in which paper they are not knowledgeable enough to review it. Once this is done for every IPC COMFy tries to create the best matching of reviewer to the submission.

Before such a matching is created COMFy currently cross checks the DBLP if there are some coauthor links which are not defined by an IPC. If some links are found the chair is notified in the suggested matching. Currently this system is redundant as IPC users are
checking their conflicts by hand. In the future this automated assignment process will be improved and the cross checks against the DBLP should replace the manual conflict settings.

7. Conclusion and Future Work

In this paper we described the new approach for a conference management system. Instead of creating a closed system the new system will allow to utilize the information of different data-sources. There is strict distinction between the core business logic of conference and the user interface defining the various use-cases. Currently the SRMv2 prototype is tested by a small conference (see figure 5(b)).

Future work will concentrate on improving the support for a semi-automatic review assignment process. For the current SRM system an experimental tool exists which calculated a so-called affinity score for the IPC members. Instead of performing a time-consuming bidding process, the IPC members upload up to 5 papers which are processed using natural language processing. These affinity scores were matched with the scores from the submitted papers in order to get an automatic distribution from IPC members to the submitted papers. Instead of the IPCs uploading their papers, the tool should harvest the papers itself from sources like the Eurographics or ACM digital library. By tweaking and improving this module the mapping should be that sufficient that the additional bidding and self entered level of expertise will be unnecessary. So the time the IPC members currently have to invest by bidding on the papers and setting the experience will be saved.

References

Full Professional Articles
This page intentionally left blank
Deduplication of Metadata Harvested from Open Archives Initiative Repositories

Piotr WENDYKIER
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, ul. Prosta 69, 00-838 Warsaw, Poland

Abstract. Open access (OA) is a way of providing unrestricted access via the Internet to peer-reviewed journal articles as well as theses, monographs and book chapters. Many open access repositories have been created in the last decade. There is also a number of registry websites that index these repositories. This article analyzes the repositories indexed by the Open Archives Initiative (OAI) organization in terms of record duplication. Based on the sample of 958 metadata files containing records modified in 2012 we provide an estimate on the number of duplicates in the entire collection of repositories indexed by OAI. In addition, this work describes several open source tools that form a generic workflow suitable for deduplication of bibliographic records.

Keywords. deduplication, record linkage, metadata, Dublin Core, open access, OAI-PMH

1. Open Archives Initiative

The Budapest Open Access Initiative [1] defines open access documents as follows: "By open access, we mean its immediate, free availability on the public internet, permitting any users to read, download, copy, distribute, print, search or link to the full text of these articles, crawl them for indexing, pass them as data to software or use them for any other lawful purpose." The number of open access repositories has grown significantly in the last decade. As a result, different organizations started creating registry websites that index these repositories. One of them is maintained by the Open Archives Initiative (OAI) - an association that develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The main goal of OAI is to enhance access to all kinds of e-print archives independent of the type of content offered as well as the economic mechanisms surrounding that content. Current OAI projects include the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Open Archives Initiative Object Reuse and Exchange (OAI-ORE). The latter one defines standards for the description and exchange of aggregations of Web resources. These aggregations may combine distributed resources with multiple media types including text, images, data, and video. In this article we focus only on the bibliographic records that can be acquired by using OAI-PMH protocol.
1.1. OAI-PMH

The OAI-PMH [2] is used to collect the metadata descriptions of the records in an archive so that services can be built using metadata from many archives. It uses XML over HTTP and the current version 2.0 was updated in 2008. OAI-PMH is based on a client–server architecture, in which harvesters request information on updated records from repositories. Requests for data can be based on a date-stamp range, and can be restricted to named sets defined by the provider. Data providers are required to provide XML metadata in a Dublin Core format [3], and may also provide it in other XML formats. The specification of OAI-PMH defines an item as a container that stores or dynamically generates metadata about a single resource in multiple formats, each of which can be harvested as records via the protocol. Each item has an identifier that is unique within the scope of the repository. The format of the unique identifier must correspond to that of the URI (Uniform Resource Identifier) syntax. A record is metadata expressed in a single format and is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item. The XML-encoding of records is organized into three parts: header (contains the unique identifier of the item and properties necessary for selective harvesting), metadata (a single manifestation of the metadata from an item), and about (an optional and repeatable container to hold data about the metadata part of the record). An example of the OAI-PMH record with metadata expressed in the Dublin Core format is shown in Listing 1.

```xml
<header>
  <identifier>oai:arXiv:cs/0112017</identifier>
  <datestamp>2002-02-28</datestamp>
  <setSpec>cs</setSpec>
  <setSpec>math</setSpec>
</header>
<metadata>
  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
           xmlns:dc="http://purl.org/dc/elements/1.1/
           xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                               http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:title>Using Structural Metadata to Localize Experience of Digital Content</dc:title>
    <dc:creator>Dushay, Naomi</dc:creator>
    <dc:description>With the increasing technical sophistication of both information consumers and providers, there is increasing demand for more meaningful experiences of digital information. We present a framework that separates digital object experience, or rendering, from digital object storage and manipulation, so the rendering can be tailored to particular communities of users.</dc:description>
    <dc:description>Comment: 23 pages including 2 appendices, 8 figures</dc:description>
    <dc:date>2001-12-14</dc:date>
    <dc:type>e-print</dc:type>
  </oai_dc:dc>
</metadata>
<about>
  <provenance>
    <originDescription harvestDate="2002-02-02T14:10:02Z" altered="true">
      <baseURL>http://the.oa.org</baseURL>
      <identifier>oai:r2:klik001</identifier>
      <datestamp>2002-01-01</datestamp>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
    </originDescription>
  </provenance>
</about>
```

Listing 1. OAI-PMH record with metadata expressed in the Dublin Core format.
OAI-PMH supports the following six types of requests (aka. verbs) [2]:

- **Identify** - retrieves information about a repository such as repositoryName, baseURL, protocolVersion, earliestDatestamp, deletedRecord, and granularity.
- **ListMetadataFormats** - retrieves the metadata formats available from a repository. Possible formats include oai_dc, olac and perseus.
- **ListSets** - retrieves the set structure of a repository, which is useful for selective harvesting.
- **GetRecord** - retrieves an individual metadata record from a repository. It is required to specify the identifier of the item from which the record is requested and the format of the metadata that should be included in the record.
- **ListRecords** - is used to harvest records from a repository that match specified criteria such as lower and upper bounds for the datestamp.
- **ListIdentifiers** - is an abbreviated form of ListRecords, retrieving only headers rather than records.

Each type of request takes additional arguments. For ListRecords, the following arguments are supported:

- **from** - an optional argument which specifies a lower bound for datestamp-based selective harvesting.
- **until** - an optional argument which specifies an upper bound for datestamp-based selective harvesting.
- **set** - an optional argument which specifies set criteria for selective harvesting.
- **resumptionToken** - an exclusive argument with a value that is the flow control token returned by a previous ListRecords request that issued an incomplete list.
- **metadataPrefix** - a required argument (unless the exclusive argument resumptionToken is used) that specifies the metadataPrefix of the format that should be included in the metadata part of the returned records.

For example, the request in the form

### 1.2. Dublin Core

The Dublin Core metadata terms [3] are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. The terms can be used to describe a full range of web resources: video, images, web pages etc. and physical resources such as books and objects like artworks. The Simple Dublin Core Metadata Element Set (DCMES) [4] consists of 15 metadata elements: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Each Dublin Core element is optional and may be repeated. The Dublin Core Metadata Initiative (DCMI), a public, not-for-profit company, has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements. The Dublin Core became ISO 15836 standard in 2006 and is a required metadata format in OAI-PMH protocol. Listing 1 shows how the Dublin Core format is used in the OAI-PMH record.
1.3. Repositories

The OAI-PMH has become widely adopted by many digital libraries, institutional repositories, and digital archives. There are several large registries of OAI-compliant repositories:

1. The Open Archives list of registered OAI repositories [5].
3. ScientificCommons.org: a worldwide service and registry [7].

This work present results only from the first registry. As of March 6, 2013, there are 1934 repositories registered in the Open Archives list including ArXiv, PubMed Central, Nature Publishing Group Metadata Repository, NASA Technical Reports Server, and CERN Document Server. Since Open Archives list stores only links to individual repositories, it is not possible to query aggregated results from multiple collections. To enable such functionality, one would have to build a digital library containing metadata from all repositories. However, to ensure uniqueness of each record in the resulting library, a deduplication operation needs to be performed. The rest of this paper describes the problem of deduplication in practice, where the database of over four million bibliographic records is analyzed.

2. Record Linkage and Deduplication

Databases play very important role in today’s world. Many industries depend on the accuracy of data stored in databases to carry out operations. Therefore, for any analysis of database resources, the uniqueness of every record is very desirable. Unfortunately, due to the lack of unique global identifier in each table, the process of linking of two or more tables may introduce duplicates. This section, after formal definition of the problem, describes the main features of a particular software package for record linkage and deduplication.

2.1. Definition

For a given sets A and B of records, a record linkage is a process of finding a partition of A×B consisting of sets M (matched), U (unmatched), and P (possibly matched) that satisfy $M = \{(a, b) \mid a = b\}$ and $U = \{(a, b) \mid a \neq b\}$. In a classical probabilistic approach to record linkage defined by Fellegi [8], a vector of similarity scores (or agreement values) is first computed for each pair. Then, the pair is classified as either a match or non-match (or possibly matched) based on an aggregate of the similarity scores. Classification methods have been widely studied [9,10,11] and the current state-of-the-art algorithms include rule-based methods that allow human experts to specify matching rules, unsupervised learning methods such as Expectation-Maximization (EM) that learns the weights or thresholds without relying on labeled data, and supervised learning methods that use labeled data to train a model, such as a decision tree, a naïve Bayesian or the Support Vector Machine (SVM). For computing similarities, various distance functions are used and studied [9,12,13,14].
Data deduplication refers to the process of identifying duplicates in a single set of records. Therefore, all the classification methods and distance functions defined for the record linkage, apply also to the problem of deduplication.

2.2. Fine-grained Record Integration and Linkage Tool

There are several open source tools for record linkage and deduplication with Link Plus [15] and Link King [16] being the most widely known examples. None of them, however, offers as many features and configuration settings as Fine-grained Record Integration and Linkage Tool (FRIL) [17,18]. FRIL (current version 2.1.5) is an open source Java software that incorporates a collection of record distance metrics, search methods, and analysis tools. In addition, FRIL provides a rich set of user-tunable parameters to enhance linking performance and accuracy [19]. FRIL’s data reconciliation feature allows merging two attributes in data source into one attribute (e.g. merging first and last names stored separately into one attribute), splitting given attribute into few columns (e.g. splitting name attribute into two attributes, first and last names) as well as trimming/replacing values in a given attribute via regular expressions. A general architecture of FRIL is shown in Figure 1.

![Figure 1. A general architecture of FRIL (taken from [18]).](image)

2.2.1. Data Sources

Data sources visible in Figure 1 provide input data for FRIL. The following data sources are supported:

- CSV file - text file where data is separated by comma, semicolon or other separator character.
- Excel file - file saved in the Microsoft Excel format.
- Fixed width columns text file - each row in that file is a separate record, and every attribute should have fixed length.
- Database - the data is pulled from a database. Currently supported: Microsoft SQL Server, PostgreSQL, MySQL, and Oracle.
Surprisingly, the only format that is supported by a result saver is a CSV file with ASCII encoding.

2.2.2. Search and Blocking Methods

Search methods, defined in the context of record linkage, refer to algorithms for determining which pairs of records to compare between the data sources A and B and the attributes on which comparisons are made. FRIL implements three types of search methods: a nested loop join, a blocking search and a sorted neighborhood method. The nested loop join performs an all-to-all comparison between A and B and is useful for small data sources. The blocking search method scans all the records in both data sources and divides them into sets called buckets based on some predefined blocking conditions. Under the assumption that no matches occur across different blocks, only records that fall into the same bucket are compared. The sorted neighborhood method first sorts records of A and B over the relevant attributes, and follows by comparing only records within fixed windows A and B of records as these windows are advanced along the input data sets.

In the context of deduplication, the equivalent of a search method is a blocking technique. FRIL implements three blocking algorithms: based on the value of blocking attribute, based on the Soundex code of a blocking attribute and based on the prefix of a blocking attribute.

2.2.3. Distance Metrics

Distance metrics are one of the most important components in the process of deduplication. They allow to provide information on how the attributes should be treated (as strings, numbers, dates) and what are allowable discrepancies between values of the attributes. Each distance function maps similarity between values of attributes onto value in the range $[0, 1]$, where 0 means “are not similar in terms of this distance function” and 1 means “are exactly the same in terms of this distance function”. FRIL implements seven distance functions:

- **Equal fields boolean distance** - returns 1 if compared values are exactly the same, otherwise it returns 0.
- **Edit distance** - it tests how many operations need to be applied to the first string so that it is converted to the other string. Possible operations include deleting a character, inserting a character or switching two characters that are next to each other. This metric requires two parameters: approve level and disapprove level. The score of edit distance function is calculated using the following formula:

$$d_e(s_1, s_2) = \begin{cases} 
0, & e(s_1, s_2) > d \cdot \max(|s_1|, |s_2|) \\
1, & e(s_1, s_2) < a \cdot \max(|s_1|, |s_2|), \\
\frac{\max(|s_1|, |s_2|) - e(s_1, s_2)}{(d-a) \max(|s_1|, |s_2|)}, & \text{otherwise}
\end{cases}$$

where $e(s_1, s_2)$ is the edit distance between two strings, $a$ is an approve level and $d$ is a disapprove level.
- **Jaro-Winkler distance** - the score of this distance is calculated as $d_w(s_1, s_2) = d_j + l \cdot p \cdot (1 - d_j)$, where $d_j$ is the Jaro-Winkler distance for strings $s_1$ and $s_2$, $l$ is the length of common prefix at the start of the string up to a maximum of 4
characters and $p$ is a constant scaling factor for how much the resulting distance $d_j$ is adjusted upwards for having common prefixes. The Jaro-Winkler distance is defined as follows:

$$d_j = \frac{1}{3} \left( \frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m - t}{m} \right),$$

where $m$ is the number of matching characters, and $t$ is the number of transpositions. Two characters from $s_1$ and $s_2$ are matching, if they are not farther than $\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1$ positions apart. The number of matching (but different sequence order) characters divided by 2 defines the number of transpositions.

- **Q-grams distance** - both input strings are first divided into q-grams (substrings of length q). Next, a count of items that do not appear in the intersection of those two sets is calculated. Configuration of this distance requires specification of q, approve level and disapprove level.
- **Soundex** - implements the Soundex algorithm [20]. Requires Soundex length, approve and disapprove levels.
- **Numeric distance** - this distance allows users to specify a range of values that will have a non-zero match score.
- **Date distance** - input values are treated as dates.

### 3. Experimental Results

The process of data analysis discussed in this work is divided into three steps: data harvesting, database creation and deduplication. The details of each step are presented in this section.

#### 3.1. Testbed

The process of data harvesting was performed on a Linux server equipped with quad-core AMD Opteron 2356 (2.30 GHz) and 16 GB RAM. The machine was running Debian GNU/Linux 6.0.3 and Oracle Java SE 6u26. Database creation and deduplication was performed on a notebook computer equipped with Intel Core i5-2430M CPU (2.40 GHz) and 8 GB RAM. The notebook was running Windows 7 64-bit operating system, MySQL 5.5 and Oracle Java SE 7u4.

#### 3.2. Data Harvesting

An XML formatted list of base URLs of registered data providers is available at [http://www.openarchives.org/Register/ListFriends](http://www.openarchives.org/Register/List_friends). To harvest the records from all providers we used an open source Java program called OAIHarvester2 [21]. For the purpose of this article, only records added in 2012 (from=2012-01-01) was harvested and the Dublin Core was used as a metadata format. This step finished within 12 hours and the harvester was able to download 1679 files out of which 717 was empty (no records were modified in 2012) and 4 files had an invalid format. Remaining 958 nonempty files, containing 4666194 records, were used in the next stage.
3.3. Database Creation

MySQL, the world’s most used open source relational database management system (RDBMS), was chosen to store the records. For the purpose of duplicates identification, the following three terms from Dublin Core metadata was chosen: title, subject and identifier. In addition, the table in the RDBMS contained the id (primary key) and the repository columns, which were not used in the process of deduplication. The process of populating the database took about an hour and it was the last step needed to prepare the input data for FRIL.

3.4. Deduplication

The process of deduplication was performed by using the latest version of FRIL. It was run three times with different distance metrics: equal fields boolean distance, Jaro-Winkler distance and edit distance. Listing 2 shows FRIL configuration file with Jaro-Winkler distance. In that file, an XML tag <params> contains information about database connection, a <row-model> tag lists all the columns that will be stored in the FRIL output files (see tags <minus-file> and <dedupe-file>) and the deduplication settings are stored in a <deduplication> tag. Three columns were selected to compute similarity measure between records and the following weights were assigned to these columns: 34 (title), 33 (subject) and 33 (identifier). For the Jaro-Winkler metric, the prefix length was set to 4 and the scaling factor was equal 0.1. The edit distance was configured with approve level = 0.0 and disapprove level = 0.1. The duplicate acceptance level was set to 100 for all metrics.

FRIL considers two records as duplicates if the sum of products of distance metric value for compared attribute between the records multiplied by the weight for the attribute is greater or equal to the duplicate acceptance level. FRIL looks for duplicates only among the records that end up in the same block. The Soundex code of the title column with a length equal to 5 was chosen as a blocking method.
Listing 2. FRIL configuration file.

Table 1 shows the results of deduplication: the number of identified duplicates, the percentage of all records and the time of computations. These values clearly demonstrate that the collection of 4666194 entries contains significant amount of duplicates. In addition, the number of duplicates depends considerably on the metric - the equal fields boolean distance and edit distance identified almost the same number of duplicates, while Jaro-Winkler metric found about 2.5 times less duplicates. This might be due to the fact that Jaro-Winkler distance is more suitable for shorter strings, such as names. Nevertheless, due to the nature of the equal fields boolean distance, we have proven that there are at least 318818 duplicates in the processed collection. Finally, the time needed for deduplication varies depending on the type of distance metric used, equal fields boolean distance was more that 190 times faster than edit distance.

Table 1. Deduplication results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Number of duplicates</th>
<th>Percentage</th>
<th>Time in hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal fields boolean</td>
<td>318818</td>
<td>6.83%</td>
<td>0.34</td>
</tr>
<tr>
<td>Jaro-Winkler</td>
<td>126721</td>
<td>2.72%</td>
<td>13.45</td>
</tr>
<tr>
<td>Edit distance</td>
<td>319099</td>
<td>6.84%</td>
<td>64.66</td>
</tr>
</tbody>
</table>

4. Conclusions and Future Work

It has been shown that the open access repositories collected by OAI contain many duplicates. Based on the experimental results, it can be estimated that duplicates pose at least 6.83% of the entire collection. Moreover, a generic workflow suitable for deduplication of bibliographic record and based on the open source tools has been described. It should be noted that these tools are suitable for processing relatively small collections of metadata, with ten million records being a reasonable limit. For very large collections containing hundreds of millions of records, the blocking methods available in FRIL would
require a machine with a lot of RAM memory and the whole process would take a long time to finish.

In the future work we plan to harvest and deduplicate all metadata records indexed by OAI. Such task is a big data challenge and needs to be solved with a dedicated software tools such as Apache Hadoop - an open source implementation of MapReduce paradigm.

References


Facilitating Access and Reuse of Research Materials: the Case of The European Library

Nuno Freire1

Abstract. The European Library provides access to research materials from the collections of Europe’s national and research libraries, representing members from 46 countries. This paper presents the current status, on-going work, and future plans of the resource dissemination services provided by The European Library, covering resources such as national bibliographies, digital collections, full text collections, its access portal and API, open linked data publication, and integration in digital humanities infrastructures. In the coming years, The European Library will work to provide the means and tools for digital humanities researchers to easily use research materials from libraries in their research activities.

Keywords. digital libraries, full text, digital humanities, portals, research infrastructures.

Introduction

The European Library provides access to research materials existing in the collections of Europe’s national and research libraries. The European Library’s partnership with libraries extends beyond the European Union, covering all the Council of Europe member states. It is a pan-European service in the widest sense.

The European Library centres its activities in the provision of services based on exploiting the centralization of pan-European bibliographic data and digital content. Its most visible service is its portal, which is designed with researchers as the target user. The portal provides value to the libraries by showcasing their collections of research materials to a wide audience, and drawing new customers to their repositories. The homepage of the portal is shown in Figure 1. The centralization of resources from European libraries also allows the provision of other services that promote the reuse of these digital resources in many contexts.

This paper starts with a presentation of The European Library, and follows with a description of the main digital resources that underlie its service. These include national bibliographies, bibliographic databases, digital objects and full text materials.

1 Corresponding Author: Senior researcher, The European Library, Europeana Foundation, Willem-Alexanderhof 5, 2509 LK The Hague, Netherlands; Email: nfreire@gmail.com
1. The Origin and History of The European Library

The European Library of today has evolved from a number of earlier projects. Its starting point was in 1997, when Project GABRIEL (Gateway and Bridge to Europe’s National Libraries) set out to establish a joint web portal of European national libraries.

At a basic level, the portal provided information about each library’s collections and access to their online public access catalogues, together with their associated web services.

Building on GABRIEL’s success, the TEL (The European Library) Project (2001–2004) went on to create a framework for access to key national and deposit collections within Europe. The project partners included the CENL, the national libraries of Finland, Germany, Italy (Florence), the Netherlands, Portugal, Slovenia, Switzerland and the United Kingdom, together with the Central Institute for the Union Catalogue of Italian Libraries (ICCU). Its achievement was to set down policies and to undertake the technical ground-work for a sustainable pan-European digital library, based on the distributed collections of the participating libraries. The overall outcome of the project was agreement to take it forward as an operational service under the auspices of CENL and hosted by the National Library of the Netherlands in The Hague, The Netherlands.
The European Library was launched as an operational service in March 2005. It aggregates the digital collections of Europe’s national and research libraries and brings together the catalogues of its partner libraries to create a union catalogue for Europe.

Between 2005 and 2007, the TEL-ME-MOR project helped to incorporate 10 more national libraries from new European Union member states as full partners of The European Library. By the beginning of 2008, a further nine national libraries had joined the service. The European Library was further enlarged by the EDL project, during which national libraries continued to join The European Library. The project also focused on multilingualism, undertook the first steps towards a European Metadata Registry and created a roadmap for potential digitization efforts in the national libraries.

The European Library provided much of the organisational structure and expertise required to launch Europeana. When the European Parliament called for Europeana to be established, The European Library was asked to apply to run the project under the eContentplus programme. Work began on a prototype in 2007 and Europeana was launched in November 2008.

Europeana now operates independently but The European Library continues to work closely with Europeana. In project Europeana Libraries, which ran from 2010 until 2012, the library aggregation model for Europeana has been established, where The European Library had the mission to become the library-domain aggregator service for Europeana. This aggregation service provided by The European Library is available for libraries across Europe.

Currently, The European Library is led by three major library associations from Europe:

- The Conference of European National Librarians\(^2\) (CENL) - The CENL foundation promotes the role of national libraries in Europe, in particular in respect of their responsibilities for maintaining the national cultural heritage and ensuring the accessibility of knowledge. Members of CENL are the national librarians of all Member States of the Council of Europe. The conference currently consists of 49 members from 46 European countries.

- Association of European Research Libraries\(^3\) (LIBER) – LIBER is the main network for research libraries in Europe, comprising more than 400 national, university and other libraries from more than 40 countries.

- Consortium of European Research Libraries\(^4\) (CERL) - CERL is a library consortium focused on the European printed heritage from the hand-press period (up to c. 1830). It aims to the sharing of resources and expertise between research libraries.

These three associations include in its members over 400 national and research libraries. For a library to be a member in The European Library, it is not required a membership in CENL, LIBER or CERL.

The European Library has launched a new web portal in 2012, aiming to open up new possibilities for academic research, and to build on The European Library’s already unique position of giving researchers free, online access to research materials from Europe’s national and research libraries.

---

\(^2\) http://web3.nl.nlib.ee/cenl/
\(^3\) http://www.libereurope.eu
\(^4\) http://www.cerl.org
2. Resources Aggregated by The European Library

This section outlines the main research resources available through The European Library. It also describes the unique characteristics of these resources, and summarises our recent and on-going activities.

2.1. Bibliographic Resources

The European Library holds several millions of bibliographic records from the national and leading research libraries of Europe. This bibliographic universe is organised into collections – groups of objects clustered around a topic or theme. The concept of collections is an extremely important one for The European Library because it represents a curatorial attempt at categorising content. Hundreds of collections are available and together they cover all the general areas of knowledge.

The National Bibliographies, one of the main bibliographic data sources in each country, are a key resource for The European Library users. Their purpose is to list every publication in a country, under the auspices of a national library or other government agency. Depending on the country, all publishers will need to send a copy of every published work to the national legal deposit, or in other countries, a national organisation will need to collect all publications. Given that the publisher domain is very heterogeneous and that thousands of publishers might exist in a country, national bibliographies are effectively the single point of reference with which to comprehensively identify all the publications in a country. In Europe, national bibliographies are typically created and maintained by national libraries. Whenever a book is published in a country, it is recorded in the corresponding national library catalogue from where the national bibliography is derived.

The European Library also offers access to collections of digital resources. These include a diverse wealth of content, such as e-theses, books, manuscripts and films, including many sizeable collections grouped around specific themes, topics or notable personalities. Examples of these are: digitised letters from prominent thinkers such as philosopher Immanuel Kant, pamphlets detailing the history of the Low Countries from 1490 to 1800, and images covering the papers, photographs and correspondence of prominent physicists such as Nobel prize winner Erwin Schrödinger. There are also digitised manuscripts created in European parts of the Ottoman Empire.

Currently, The European Library holds approximately 112 million bibliographic records in its centralized repository. This number is constantly increasing, as part of its data aggregation role in the Europeana context, and more national and research libraries’ catalogues are included in the centralized repository. By the end of 2013, the total bibliographic universe of The European Library is expected to be approximately 200 million records.

2.2. Textual Resources

The European Library hosts a growing centralized index of textual resources. It currently contains over 24 million pages of full-text content, originating from 14 national libraries.

These textual resources were created mostly from Optical Character Recognition performed during digitization projects at the national libraries. The quality of the text
varies, depending on the quality of the original material, and the use of special fonts, some of which remain challenging for even modern readers.

That said, the size of the full-text collection, and the fact that it is relatively heterogeneous in terms of types of materials, languages and publication period, makes it a relevant resource for digital humanities research. An overview of the available resources is provided in Table 1.

Table 1. Textual resources currently indexed at The European Library.

<table>
<thead>
<tr>
<th>Country of origin</th>
<th>Material type</th>
<th>Pages</th>
<th>Temporal coverage</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>Newspapers, governmental material</td>
<td>534,000</td>
<td>1862 – 1925</td>
<td>German</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>Books, newspapers</td>
<td>2,579,511</td>
<td>1800 – 1989</td>
<td>Czech, German</td>
</tr>
<tr>
<td>Estonia</td>
<td>Newspapers, journals</td>
<td>713,933</td>
<td>1821 – 1940</td>
<td>Estonian</td>
</tr>
<tr>
<td>France</td>
<td>Books, periodicals</td>
<td>8,242,908</td>
<td>1650 – 1930</td>
<td>French (some others)</td>
</tr>
<tr>
<td>Hungary</td>
<td>Periodicals, newspapers, books, journals, monographs, pamphlets</td>
<td>237,914</td>
<td>1590 – 1992</td>
<td>Hungarian, Latin, English, German</td>
</tr>
<tr>
<td>Iceland</td>
<td>Newspapers, journals</td>
<td>5,727,149</td>
<td>1773 – 2002</td>
<td>Icelandic, Faroese, Greenlandian</td>
</tr>
<tr>
<td>Latvia</td>
<td>Newspapers, books</td>
<td>195,075</td>
<td>1900 – 1952</td>
<td>German, Latvian</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Newspapers</td>
<td>125,477</td>
<td>1904 – 1940</td>
<td>Lithuanian</td>
</tr>
<tr>
<td>Norway</td>
<td>Books, journals</td>
<td>1,600,000</td>
<td>By authors dead for more that 70 years</td>
<td>Norwegian (others)</td>
</tr>
<tr>
<td>Poland</td>
<td>Newspapers, books</td>
<td>436,198</td>
<td>Before 1939</td>
<td>Polish, German, Czech, Ukrainian, Belarusian, Yiddish</td>
</tr>
<tr>
<td>Slovakia</td>
<td>Newspapers</td>
<td>185,000</td>
<td>Before 1918</td>
<td>Slovak, Hungarian, German</td>
</tr>
<tr>
<td>Slovenia</td>
<td>Newspapers, books, journals</td>
<td>328,502</td>
<td>1500 – 1945</td>
<td>Slovenian</td>
</tr>
<tr>
<td>Spain</td>
<td>Newspapers, books</td>
<td>3,033,525</td>
<td>17th – 19th Century</td>
<td>Spanish</td>
</tr>
<tr>
<td>Sweden</td>
<td>Newspapers, books, journals, printed ephemera</td>
<td>253,653</td>
<td>Until the 20th century</td>
<td>Swedish</td>
</tr>
</tbody>
</table>
These textual resources will be expanded during 2013, thanks to the Europeana Newspapers project. In this project, a group of 17 European institutions are working to provide more than 18 million newspaper pages for The European Library and Europeana. Each library participating in the project will distribute digitized newspapers and full-texts that are in the public domain and free of any legal restrictions.

3. Resource Dissemination Services

Although The European Library service was founded as a traditional library service, focused on making the library resources available through traditional online catalogue features, it is now aiming to facilitate the use of these resources in order to improve existing services, enable the development of new ones, or provide a good setting for research. This section presents the main services and activities undertaken by The European Library.

3.1. The European Library Portal

The European Library’s portal main focus is to provide access to the collections of the European national and research libraries. It provides a web interface for searching and browsing through libraries, collections, bibliographic records, digital objects and full text contents. A new version of the portal was launched in May 2012, which included a complete functional redesign, focused on a shift towards functionalities for digital humanities researchers.

While previous versions of the portal were based on the distributed search paradigm, this new version is based on the centralization of the digital resources, which enables much functionality that nowadays web users expect in resource discovery portals.

Data mining technologies are being applied for automatic linkage of bibliographic data across countries and libraries (some examples are shown in Figure 2). On-going work is focusing on the linkage of the main entities used to describe bibliographic resources: persons, organizations, locations, historical periods, and subjects, in order to enable better functionality in the portal. This on-going work, once completed, will allow the exploration of an author’s work across time and space. Users will be able to locate different editions of a work throughout Europe, browse all catalogues with a unified subject system, and browse across historical periods.

Access to the bibliographic resources under a unified subject system is also a core functionality of the portal. Subject information is frequently used in bibliographic data and very relevant for the portal functionality, given the multilingual nature of The European Library. However, unifying subject systems is a great challenge because of the diversity of languages and knowledge organization systems in use across European libraries. Unified subject access is currently based in the results of the project Multilingual Access to Subjects (MACS). The MACS project has produced manual

---

5 http://www.europeana-newspapers.eu/
6 http://www.d-nb.de/eng/wir/projekte/macs.htm
alignments between three major systems: the Library of Congress Subject Heading\(^7\) (LCSH), the Répertoire d’autorité-matière encyclopédique et alphabétique unifié\(^8\) (RAMEAU) and Schlagwortnormdatei\(^9\) (SWD), covering the English, French and German languages. Ongoing work is targeting wider coverage of subject systems, by exploring the alignment of language independent subject classification systems (such as Dewey Decimal Classification\(^10\) and Universal Decimal Classification\(^11\)), and by researching semi-automatic subject alignment techniques.

Although subject information in our datasets is based on the above mentioned library subject systems, the unification of subject access is being implemented based on the R&D subject indexing scheme of CERIF.

![Figure 2. Mining bibliographic data at The European Library’s portal.](image)

---

7 http://id.loc.gov/authorities/
8 http://rameau.bnf.fr/
9 http://www.d-nb.de/standardisierung/normdateien/swd.htm
10 http://dewey.info/
11 http://www.udcc.org/about.htm
3.2. The European Library API

The European Library OpenSearch\(^\text{12}\) API allows the development of applications by third parties, which allow the search and display of The European Library’s collections, bibliographic data and previews. Along with standard keyword searching, the API also offers faceted searching. Currently, the OpenSearch API is available for non-commercial use to all members of The European Library.

We are also working towards making the resources hosted at The European Library available through third party applications in the research domain, such as Mendeley\(^\text{13}\) and the Summon discovery service\(^\text{14}\).

3.3. Aggregation for Europeana

Europeana\(^\text{15}\) provides access to digital content from libraries, museums, archives and audio-visual collections across Europe. Data about its digital contents follows the Europeana Data Model [1] (EDM) which is an open, cross-domain, semantic web-based framework for structuring the data about the digital content that Europeana ingests, manages, publishes for end-user access, and also publishes as open data. This data model allows Europeana to keep the richness of domain-level metadata and facilitates the participation of Europeana in the semantic web.

The European Library conducts activities within the library domain, involving EDM, in order to promote the representation of library specific information in EDM alongside authoritative and curated information from other domains. Of particular relevance is analysis of how bibliographic data (e.g. MARC21, UNIMARC, MODS, METS and Dublin Core-based application profiles) can be aligned to EDM, since in its role as aggregator for Europeana, The European Library supports all these bibliographic data formats for conversion and delivery to Europeana, according to EDM.

3.4. Intellectual Property Rights Infrastructures

The European Library’s bibliographic data is enabling an innovative approach towards intellectual property rights clearance processes for supporting mass digitization projects. In the ARROW rights infrastructure\(^\text{16}\) (Accessible Registries of Rights Information and Orphan Works), a framework is being established to assist in determining the rights status of works.

ARROW provides automated ways to clear the rights of the books to be digitised. This rights clearance process is time consuming, since a library has to go through the following steps for each book:

- To identify the underlying work incorporated in the book to be digitised;
- To find out if the underlying work is in the public domain or in copyright, and whether it is an orphan work or out-of-print;

---

\(^{12}\) http://www.opensearch.org/
\(^{13}\) http://www.mendeley.com
\(^{14}\) http://www.proquest.com/
\(^{15}\) http://www.europeana.eu
\(^{16}\) http://www.arrow-net.eu/
To clearly describe the use that is requested for the book, such as digitisation for preservation, electronic document delivery, etc.;
- To identify the rights holder(s) or their agent, such as a collecting society;
- To seek the appropriate permission, if necessary.

The successful completion of this process depends on the availability of existing bibliographic and rights data, and national bibliographies are effectively the single point of reference with which to identify all the publications of an intellectual work. In this infrastructure, The European Library provides web services for the processing and use of national bibliographies during rights clearance processes. The service from The European Library is thereby a bibliographic record clustering service for national bibliographies. The service provides a set of related publications from the national bibliographies for a deeper analysis in a rights clearance process.

3.5. Linked Open Data

The European Library is currently working in its open linked data prototype. This prototype aims to make openly available the results of the data mining processing applied while aggregating library catalogues, digital libraries and open repositories. This process aims to establish links between entities defined in the Functional Requirements for Bibliographic Records [2] (FRBR) such as:

- Linking between the FRBR group 1 entities, particularly work and manifestation, across libraries.
- Linking between FRBR expressions within the same country (a result of the processing for the ARROW rights infrastructure)
- Linking of FRBR group 2 entities person and corporate body, between the aggregated bibliographic data and VIAF [3] (when they are not already linked at the source).
- Linking of the FRBR group 3 entity concept, across the aggregated bibliographic data sources, by exploiting the results of the MACS project.
- Linking of the FRBR group 3 entity place between the aggregated bibliographic data and Geonames [4] (a geographic ontology available as open data).

The European library’s open data prototype aims to make this data enrichment available in a way that complements the ongoing activities of libraries with open data.

3.6. Digital Humanities Research Infrastructures

In order to provide the means and tools for digital humanities researchers to exploit the research materials held by libraries, a new research infrastructure, Europeana Research, will be created by extending the currently existing portal of The European Library.

Early stages of the project will analyse how academic users locate data and how they perceive the value of the content within Europeana. This analysis will be the basis of the content strategy of the Europeana Research platform, and will also provide the understanding of scholarly workflows which will be supported by Europeana Research. This analysis will be carried out jointly with the DARIAH network of arts and humanities researchers, and with the Council of European Social Science Data Archives (CESSDA). Academics working in these domains will be the most fertile
exploiters of libraries research materials, and therefore they have key roles to play in shaping the future of Europeana Research.

This analysis will allow the identification of tools that allow researchers to manipulate and exploit research materials in innovative ways. The project will therefore develop a suite of tools that allows scholars to interact with the content that they require from Europeana Research. The areas to be approached are:

- Accessing and analysing big data - permitting scholars to download, manipulate and analyse large data sets.
- Annotation - allowing researchers to annotate documents and to share these annotations
- Transcription - allowing users to transcribe and interpret documents
- Discovery and access - ensuring that services are tailored so that research material is discoverable by the scholarly community, possibly with integration in other research infrastructures in the field of digital humanities.

The creation of statistical datasets, resulting from data analysis and mining conducted by The European Library, is also being considered. Although no concrete plans are ready at this time, we expect these kinds of data sets to emerge as a result of our involvement with the digital humanities research community.

4. Conclusion

This paper presented The European Library and the main digital resources that underlie its services. The European Library provides services that add value to resources made available by libraries, by exploring the possibilities given by centralization of data and resources.

The European Library is entering a new phase as national and research libraries are coming together in a partnership to bring libraries’ research collections to a worldwide audience. The European Library aims to support libraries to play a pivotal role in Europe’s research agenda, supporting the growth strategy of the European Commission, which highlights higher education as key to fostering innovation, employment and long-term economic prospects. These policy developments present exciting opportunities for national and research libraries to work with Europeana to develop a research platform to support innovation and growth.

References

Exploring FP7 Funded PLOS Publications

Najko JAHN a,1, Martin FENNER b, c and Jochen SCHIRRWAGEN a

a Bielefeld University Library, Bielefeld University, Germany
b Hannover Medical School, Germany
c Technical Lead PLOS Article Level Metrics, Public Library of Science, USA

Abstract. This case study explores alternative science metrics on grant-supported research publications. The study is based on plosOpenR, a software package for the statistical computing environment R. plosOpenR facilitates access to the application programming interfaces (API) provided by Open Access publisher Public Library of Science (PLOS) and OpenAIRE – Open Access Infrastructure for Research in Europe.

We report 1,166 PLOS articles that acknowledge grant support from 624 different research projects funded by the European Union’s 7th Framework Programme (FP7). plosOpenR allows the exploration of PLOS Article-Level Metrics (PLOS ALM), including citations, usage and social media events as well as collaboration patterns on these articles. Findings reveal the potential of reusing data, that are made openly and automatically available by publishers, funders and the repository community.

Keywords. Research evaluation, Article-Level Metrics, Statistical Computing, R, PLOS, Open Access, OpenAIRE, 7th Framework Programme

1. Introduction and Motivation

With the growing number of openly available research services, new opportunities arise for measuring performance and impact of research publications. The quantitative study of scholarly communication no longer solely depends on traditional citation analysis, but is complemented by usage and social media data. Publishers, funders and the repository community are likewise seeking for ways to provide relevant data for the ongoing work on alternative metrics.

Since its launch, the Open Access publisher Public Library of Science (PLOS)2 has been a strong advocate of alternative ways to publish and measure research. On every article, PLOS displays a suite of indicators including citations, information on usage and social media activity.3 In November 2012, PLOS also started the Altmetrics collection inviting research articles on alternative science metrics [1].

1 Corresponding Author: Bielefeld University Library, Bielefeld University, Universitaetsstrasse 25, 33602 Bielefeld, Germany; E-mail: najko.jahn@uni-bielefeld.de.
2 Public Library of Science (PLOS), http://www.plos.org/
3 PLOS Article-Level-Metrics (ALM), http://article-level-metrics.plos.org/
It is promising to apply PLOS Article-Level Metrics (PLOS ALM) in broader contexts, as for instance on a particular funding programme. OpenAIRE – Open Access Infrastructure for Research in Europe⁴ gives access to results funded by the European Union’s 7th Framework Programme (FP7) in general and FP7 grants subject to the Open Access pilot of the European Commission under the Special Clause 39 (SC39) in particular. OpenAIRE features a dedicated work package for exploiting usage metrics on Open Access publications as a supplement to conventional citation analysis. To this end, OpenAIRE has started to aggregate data from federated usage data providers.⁵

In this case study, we stress the potential of PLOS ALM information on exploring FP7 grant-supported research publications. Moreover, our considerations reflect a growing community of scientists that collaborate on statistical tools to access and analyse research output[3,4]. The case study, therefore, derives from a set of tools for the statistical computing environment R⁵ – plosOpenR.⁵

After introducing the APIs and existing software solutions used in the next section, we demonstrate plosOpenR on the basis of 1,166 PLOS research articles acknowledging 624 FP7 funded projects. Finally, we discuss the potential benefits and limits from the perspectives of research infrastructure and quantitative science studies.

The dataset of our case study is available at doi:10.5281/ZENODO.1239

2. Background and Data

PLOS offers two public available APIs that plosOpenR uses. The PLOS Search API⁶ gives developers access to the fulltext-corpus of all PLOS articles published. The search fields correspond to article sections. For our study on information about FP7 funded research published in PLOS, we mainly rely on the fields listed in Table 1.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>DOI (Digital Object Identifier)</td>
</tr>
<tr>
<td>financial_disclosure</td>
<td>Funding acknowledgement (free text)</td>
</tr>
<tr>
<td>affiliate</td>
<td>Affiliation of the authors (free text)</td>
</tr>
</tbody>
</table>

The PLOS ALM API is used to retrieve metrics on PLOS articles.⁷ PLOS releases the underlying software under an Open Source licence.⁸ For our work, we analysed the following PLOS ALM providers listed in Table 2.

OpenAIRE exposes FP7 funding information via its OAI-PMH interface⁹ which reflects the OpenAIRE data model[6]. Detailed project information are listed in the set project. Fields that we have used to identify and contextualize projects in corre-

⁴ OpenAIRE, http://www.openaire.eu/
⁵ plosOpenR repository on GitHub, https://github.com/articlemetrics/plosOpenR
⁷ PLOS ALM FAQ, http://api.plos.org/alm/faq/
⁸ ALM software repository: https://github.com/articlemetrics/alm/
⁹ OpenAIRE BaseURL, http://api.openaire.research-infrastructures.eu:8280/is/mvc/openaireOAI/oai.do
Table 2. PLOS ALM source fields examined

<table>
<thead>
<tr>
<th>Family / Provider</th>
<th>ALM-Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLOS</td>
<td>counter</td>
<td>HTML article view, pdf and XML downloads (COUNTER 3)</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>pmc</td>
<td>HTML article view, pdf and XML downloads</td>
</tr>
<tr>
<td>Citations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PubMed Central</td>
<td>pubmed</td>
<td>Times cited for an article from PubMed Central</td>
</tr>
<tr>
<td>CrossRef</td>
<td>crossref</td>
<td>Times cited for an article from CrossRef</td>
</tr>
<tr>
<td>Scopus</td>
<td>scopus</td>
<td>Times cited an article from Scopus</td>
</tr>
<tr>
<td>Social media events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>twitter</td>
<td>Tweets for an article</td>
</tr>
<tr>
<td>Facebook</td>
<td>facebook</td>
<td>The number of Facebook Likes for an article</td>
</tr>
<tr>
<td>Mendeley</td>
<td>mendeley</td>
<td>The number of times a user has bookmarked an article in Mendeley</td>
</tr>
<tr>
<td>CiteULike</td>
<td>citeulike</td>
<td>The number of times a user has bookmarked an article in CiteULike</td>
</tr>
<tr>
<td>PLOS Comments</td>
<td>ploscomments</td>
<td>The number of times a user has comment an article on PLOS</td>
</tr>
</tbody>
</table>

spondence with acknowledgement in PLOS articles are GrantID, acronym, title, call id, fundedby, fundedhow, ac39.

plosOpenR follows three steps to explore PLOS ALM on FP7 grant-supported research publications:

- Retrieve a set of articles through the Search API
- Collect the metrics for these articles
- Visualize the metrics

For this purpose, plosOpenR reuses already existing tools to query and analyse PLOS research output that belong to the rplos package. rplos is developed by rOpenSci, a collaborative effort to provide R-based applications for facilitating Open Science.10

After querying the PLOS APIs, plosOpenR transforms the so retrieved JSON and XML outputs into data.frame structures in order to allow easier statistical analysis within R.

PLOS ALM are much easier to understand through visualizations. R provides powerful graphic devices often used in statistics. plosOpenR demonstrates different visualisation techniques which are documented on the PLOS API webpage11. For our case study, we focus on alternative scatterplots to explore ALM distributions, on network visualisa-

---

10 rOpenSci: http://ropensci.org/
tions to examine collaboration patterns[7] and on choropleth maps displaying author’s country of affiliation. For the latter, plosOpenR uses the Thematic Mapping API.12

To allow a broader query and more reliable match of FP7 funding acknowledgement visible in PLOS articles, a processing step outside of R applies OpenAIRE’s text mining, rule-based named entity recognition approach to identify FP7 project references.[8]

3. Results

3.1. FP7 Contribution in the PLOS Domain

PLOS gives access to grant information in the financial disclosure section. The openly available PLOS Search API allows specific queries of this section. On 19 July 2012, we obtained 2,562 candidate publications after querying the search field financial_disclosure:

\[
((europ* \text{ AND (union OR commission)) OR fp7}) \text{ OR ((seventh OR 7th) AND framework}) \text{ OR (ERC OR (European Research Council)) OR ((EU OR EC) AND project)}
\]

In total, we matched 1,166 PLOS articles that referenced at least one FP7 research project. The FP7 acknowledgement by PLOS journal and publishing year show a moderate growth in most journals, but a strong growth in PLOS ONE (Table 3). This journal represented 77.78% of FP7-supported research publications.

Table 3. FP7 funding acknowledgement in PLOS journals 2008–2012 (*until 19 July 2012)

<table>
<thead>
<tr>
<th>Journal / Publishing Year</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012*</th>
<th>(\Sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLOS ONE</td>
<td>8</td>
<td>36</td>
<td>132</td>
<td>358</td>
<td>335</td>
<td>869</td>
</tr>
<tr>
<td>PLOS Pathogens</td>
<td>1</td>
<td>8</td>
<td>15</td>
<td>41</td>
<td>26</td>
<td>91</td>
</tr>
<tr>
<td>PLOS Genetics</td>
<td>10</td>
<td>17</td>
<td>20</td>
<td>26</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>PLOS Computational Biology</td>
<td>1</td>
<td>10</td>
<td>16</td>
<td>21</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>PLOS Biology</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>PLOS Neglected Tropical Diseases</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>PLOS Medicine</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>(\Sigma)</td>
<td>11</td>
<td>70</td>
<td>194</td>
<td>462</td>
<td>429</td>
<td>1166</td>
</tr>
</tbody>
</table>

On this basis, we calculated the compound annual growth rate for PLOS ONE as being 215.35 % over the two-year period from 2009 to 2011. This number is consistent with the overall fast growth of the journal.

We identified 624 FP7 projects that were acknowledged in PLOS-journal articles. Table 4 presents the distribution over projects by its number of publications in PLOS.

The figures indicate a positively skewed distribution of FP7 contributions in PLOS with 57.53 % of the FP7 projects referenced once, while 1.92 % published more than 8 times with PLOS. With 30 contributions, the FP7 research project European Network

Table 4. PLOS contributions by EC funded research projects

<table>
<thead>
<tr>
<th>PLOS per FP7</th>
<th>Frequency</th>
<th>Relative Frequency (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>359</td>
<td>57.53</td>
</tr>
<tr>
<td>2</td>
<td>129</td>
<td>20.68</td>
</tr>
<tr>
<td>3 – 8</td>
<td>124</td>
<td>19.87</td>
</tr>
<tr>
<td>9 – 30</td>
<td>12</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>624</td>
<td>100</td>
</tr>
</tbody>
</table>

for Genetic and Genomic Epidemiology (ENGAGE)\(^{13}\) published the highest number of PLOS articles.

The OpenAIRE OAI-PMH interface provides more detailed access to FP7 funding information. At the time of our study, 17,736 FP7 research projects were exposed which were distributed over 23 funding programmes. On this basis, we determined the visibility of FP7 funding programmes in the PLOS domain (Table 5). Similar to the distribution over FP7 research projects, FP7 grant acknowledgements were unequally distributed over funding programmes which coheres PLOS’ focus on biomedical research and related fields: We found that 27.96 % of the projects funded within the research programme Health Research (HEALTH) published at least once in PLOS, but we could not detect references to eight funding programmes, e.g. Transport (including Aeronautics) (TPT).

We also took projects under the SC39 clause into account and revealed that the proportion of SC39 funded research in the PLOS domain (94 out of 624 FP7 projects, 7.34 %) was higher than the FP7 funding scheme share (530 out of 17,736 FP7 projects, 3.22 %).

3.2. Article-Level Metrics

The PLOS ALM API provides information on citations, usage and dedicated social media events. Figure 1 shows the coverage of FP7 articles by ALM source field until 3 September 2012.

For every PLOS contribution in our sample, we were able to collect usage data from both the PLOS journal website (counter) and from PubMed Central (pmc). However, coverage within the ALM categories of citation and social media events is more heterogeneous: between 43 % (pubmed) and 63 % (crossref) of the articles were cited. Social media services mentioned between 8 % (comments on PLOS articles) and 81 % (Mendeley readerships) articles granting FP7 support. Note that the collection of Twitter mentions within PLOS ALM started on June 1st, 2012.

Data from the PLOS ALM API furthermore allows us to compare the occurrences of ALM event types for every day since publication. Figure 2 depicts article age (in days since publication) and total views on the PLOS website. As a third variable we compared Scopus citations and Facebook shares received for each FP7 funded research article in PLOS journals (mapped as point size). Citation rates are time-dependent with few citations in the first 12 months of publication, and the article set contains many articles published in 2012. With 38,361 total views, the following article resulting from FP7 funded

\(^{13}\) ENGAGE project information, http://cordis.europa.eu/projects/201413
Table 5. Comparing the proportion of FP7 funded projects and their acknowledgement in PLOS articles by funding programmes

<table>
<thead>
<tr>
<th>EC Funding Programme</th>
<th>Projects funded</th>
<th>Projects acknowledged in PLOS</th>
<th>Ratio (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEALTH</td>
<td>769</td>
<td>215</td>
<td>27.96</td>
</tr>
<tr>
<td>KBBE</td>
<td>421</td>
<td>46</td>
<td>10.93</td>
</tr>
<tr>
<td>GA</td>
<td>25</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>INFRA</td>
<td>311</td>
<td>16</td>
<td>5.14</td>
</tr>
<tr>
<td>ENV</td>
<td>406</td>
<td>20</td>
<td>4.93</td>
</tr>
<tr>
<td>REGPOT</td>
<td>169</td>
<td>8</td>
<td>4.73</td>
</tr>
<tr>
<td>ERC</td>
<td>2909</td>
<td>122</td>
<td>4.19</td>
</tr>
<tr>
<td>ICT</td>
<td>1731</td>
<td>67</td>
<td>3.87</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>7878</td>
<td>115</td>
<td>1.46</td>
</tr>
<tr>
<td>NMP</td>
<td>584</td>
<td>8</td>
<td>1.37</td>
</tr>
<tr>
<td>Fission</td>
<td>101</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>SIS</td>
<td>142</td>
<td>1</td>
<td>0.70</td>
</tr>
<tr>
<td>SEC</td>
<td>198</td>
<td>1</td>
<td>0.51</td>
</tr>
<tr>
<td>ENERGY</td>
<td>303</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>SME</td>
<td>694</td>
<td>1</td>
<td>0.14</td>
</tr>
<tr>
<td>Fusion</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COH</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INCO</td>
<td>126</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CIP-EIP</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TPT</td>
<td>521</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>REGIONS</td>
<td>65</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPA</td>
<td>162</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SSH</td>
<td>180</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17736</strong></td>
<td><strong>624</strong></td>
<td><strong>3.52</strong></td>
</tr>
</tbody>
</table>

**Forecasting Financial Crisis** (FOC-II)\textsuperscript{14} was most represented in our sample:


However, while the article had received the highest number of Facebook shares in our study, the paper was only ranked in the 87th percentile according to Scopus citation count (10 times cited). A possible explanation gives a blog post from the FOC project. The project members describe that the article has gained broad media attention. However, the authors claim that media have misinterpreted their findings.\textsuperscript{15}

\textsuperscript{14} FOC-II project information. http://cordis.europa.eu/projects/255987
3.3. Collaboration Patterns

Our dataset can also be used to explore collaborations patterns between FP7 research projects by joint PLOS publications. We found that 9.52% of PLOS publications under consideration acknowledged more than one FP7 project (Table 6). From 624 FP7 projects identified, 26.28% were acknowledged together with another FP7 projects. Therefore, the relationships between projects and articles was furthermore explored as a scientific collaboration network.

Figure 4 visualises the links between FP7 research projects that contributed to at least one PLOS publication together. In this figure, edge width represents the number of joint PLOS contributions. In total, we detected 57 components that consist of 164 projects. The FP7 project ENGAGE is most frequently represented in our sample again,
Table 6. FP7 funding acknowledgement per PLOS article

<table>
<thead>
<tr>
<th>FP7 Projects per paper</th>
<th>Frequency</th>
<th>Relative Frequency (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,055</td>
<td>90.48</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>7.98</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>1.20</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.34</td>
</tr>
<tr>
<td>∑</td>
<td>1,166</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 3. Collaboration network of FP7 projects in PLOS journals. Edge width show the number of joint articles of a collaborating institutional pair, vertex size represent the degree centrality, i.e. the number of articles in PLOS.

counting for both the most PLOS contributions and the highest number of direct links in the network (20).

Lastly, we used the PLOS Search API to author affiliations. In total, we obtained 6,090 author addresses distributed over 1,166 publications. 6,049 correctly formatted country names could be extracted that are distributed over 95 countries (Figure 4). Affiliations listed originated most frequently from United Kingdom (12.25 %), Germany (11.77 %) and the United States (11.69 %).

If compared to World Bank regions,16 more than every third affiliation listed originated from the region of Western Europe. Regarding the distribution of PLOS publications over FP7 projects and countries, this distribution is positively skewed with the regions of Western Europe, Northern Europe, Southern Europe and Northern America accounting for 92.97 % of all affiliations detected.

4. Discussion & Conclusion

This case study presents first findings of plosOpenR, a software package for the statistical computing environment R. plosOpenR allows the exploration of grant-supported research publication in PLOS journals. We report 1,166 articles that acknowledge 624 different research projects funded by the European Union 7th Framework Programme (FP7). Additionally, our case study presents metrics such as citations, usage and social media activity as well as collaboration structures between FP7 projects visible in joint publications.

From a research infrastructure and services point of view, our results highlight the importance of openly available research services. plosOpenR combines data sources from the Open Access publisher PLOS and OpenAIRE as well as reuses existing software packages from rOpenSci. While PLOS provides detailed information on articles, OpenAIRE enables a distinct view on FP7 funded research published in PLOS journals by exposing the relevant project information. Furthermore, we demonstrated the aggregation of alternative science metrics by the PLOS ALM API. For all research contributions under investigation, we were able to retrieve usage data on a daily basis from the PLOS journal website and the disciplinary repository PubMed Central hosted by NIH (PMC).

When discussing the results in the light of quantitative science studies on performance and impact of research publications, it has to be noted that our study is limited in various ways. Firstly, we were only able to examine research contributions published in PLOS journals until 19 July 2012. The continuing increase in FP7 funded PLOS papers from January to June 2012 as well as the duration of the Seventh Framework Programme suggests that potentially more FP7 funded research publications in PLOS journals are to be expected in future. Secondly, the extreme positive skewness of most distribution
under considerations demands careful analysis and interpretation. Especially, it has to be noted that our findings only partially cover all FP7 funding projects and programmes due to the disciplinary scope of PLOS.

Particular care needs to be taken if future studies rank research articles according to the different metrics in use and develop comparative indicators that rely on these data. For instance, our exploration of Scopus citation counts in comparison with the social media event type Facebook shares on the article level revealed that public media attention has effects on analysing and interpreting research publications. In addition, whereas the majority of usage data and social web activity happens in the days and months after publication, citation data are accumulating much more slowly. The set of FP7 funded PLOS articles that we identified should therefore be reanalyzed at least two years after the last paper in the set has been published. However, with data sources and visualization methods suggested, plosOpenR provides tools for easy on-time exploration of PLOS ALM in order to identify irregular patterns and motivate qualitative investigation.

Future work and studies on PLOS ALM will focus on main problem areas [9]. With the evolving European federation of usage-data providers, OpenAIRE has the potential to provide additional information about usage events and might complement PLOS ALM as PMC already does.

Acknowledgements

We thank Harry Dimitropoulos for applying his text mining/rule-based named entity recognition approach on our sample. Jochen Schirrwagen gratefully acknowledges support from the European Commission (FP7-INFRA-2007-1.2.1, Grant Agreement no. 246686).

References

Extended Abstracts
This page intentionally left blank
Making Sense of a Flat List of Terms into Linked Open Data SKOS Vocabularies

Helder Noel FIRMINO and Ana Alice BAPTISTA

CentroAlgoritmi, University of Minho, Guimarães, Portugal

Abstract. Problem and Objectives. ANACOM (Portuguese National Authority on Communications) is the regulator, supervisor and representative of the communications sector in Portugal. Through the years, ANACOM has been developing a kind of a flat controlled vocabulary: a list of controlled terms, without any structure and without specific a priori relations to the DC properties in use. One of the requirements of the project now in hands is to organize that list into one or more controlled vocabularies, relate them with other LOD vocabularies, encode them in SKOS and make them open in at least two languages (Portuguese and English).

Methodological Approach. We will adopt a hybrid approach, combining the best features of two methodologies: Ontology Development 101(1) and the "Process and Methodology for Developing Core Vocabularies" used by the European Programme ISA (2). We will use qualitative research techniques, such as interviews and focus groups to analyse and validate the relations between the terms.

Expected Results. This article presents a work in progress comprised of the whole process starting on the organization of the terms and ending in the first version of the controlled vocabularies encoded in SKOS.

Keywords. Linked open data, SKOS, Controlled Vocabularies

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

1 Corresponding Author: Helder Noel Firmino, CentroAlgoritmi, University of Minho, Portugal; E-mail: pg19785@alunos.uminho.pt.
Does Openness and Open Access Policy Relate to the Success of Universities

Pekka OLSBO

Jyväskylä University Library

Abstract. The universities of the world are ranked by several different research institutions. In this study we will take a closer look at the Ranking Web of Universities (RWU) ranking, because it claims to be more extensive than the others and because its focus is on academic web presence of the universities. RWU has analyzed over 21 000 universities and the actual ranking covers 12 000 universities. This paper examines the internet visibility of the University of Jyväskylä and eight European countries and how the openness of universities has developed during last two editions of the Ranking Web of Universities. The Finnish Academy published a report The state of scientific research in Finland 2012 in October 2012. In this report The Academy is analyzing the relative citation impact of eight European countries. Analysis shows the good values of Switzerland, Netherlands and Denmark. If we analyze the statistics in the RWU and the Ranking Web of Repositories we might find connections between the success of universities, ranking in openness and repositories in Switzerland, Netherlands and Denmark. From the comparison of eight European countries we can find out that at least in Netherlands and Sweden there is a deep connection between the ranking of universities, repositories and the openness of the universities. But the most interesting developments can be seen, if we look at the trends of these numbers and figures. If we compare the top 5 universities of these countries we can see, that Finland, Denmark and Norway have improved their placing in openness substantially compared to Switzerland and even the whole RWU. These same countries seem to be in their way up also if we look at the development of relative citation impact in recent years. One explaining factor could be the Open Access policy of these countries and universities. Is it due to proper Open Access policy of the country, coincidence or unreliability of the methodology of Ranking Web of Universities is hard to say yet.

Keywords. Internet Visibility, Relative Citation Impact, Institutional Repositories

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

1 Corresponding Author: Pekka Olsbo, Jyväskylä University Library, Finland; E-mail: pekka.olsbo@jyu.fi.
Usage Data of an Open Access e-Journal in a Digital Repository

Teresa S. WELSH* and Stacy CREELb,1
a The University of Southern Mississippi
b The University of Southern Mississippi

Abstract. The University of Southern Mississippi’s School of Library and Information Science (SLIS) publishes SLIS Connecting in the university’s digital repository, Aquila Digital Community (http://aquila.usm.edu/), which is hosted through Digital Commons. The purpose of SLIS Connecting is “to share news, information and research with future students, current students, alumni, faculty, and the general population through selected faculty publications, invited student publications, refereed publications, and through regular columns” [1]. The first issue was electronically published in February 2012 and the second in October 2012. The third issue was published in February 2013 and contains the first paper submitted from an author not affiliated with SLIS. SLIS Connecting is currently indexed in Google Search and in Google Scholar. From the first issue of SLIS Connecting in February 2012 through March 2013, there were over 7,000 page views. Eighty-three percent were from U.S. While the patterns in the United States align closely to the student and alumni distribution as expected; the international pattern at first glance seems somewhat unexpected, but there are several possible reasons for the unexpected international reach. This extended abstract presents the spatial analysis of usage data of SLIS Connecting in the United States and abroad.


Keywords. Digital Repositories, Geographic Distribution

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

1 Stacy L. Creel, Assistant Professor, 118 College Drive #5146, Hattiesburg, MS 39406 Stacy.Creel@usm.edu.
Publisher Interest Towards a Role for Journals in Data-Sharing: The Findings of the JoRD Project

Marianne BAMKIN¹,², Jane SMITH¹, Paul STURGES³, Azhar HUSSEIN¹ and Bill HUBBARD¹

¹Centre for Research Communication, University of Nottingham
²Emeritus Professor, Loughborough University

Abstract. The sharing of the data generated by research projects is increasingly being recognised as an academic priority by funders and researchers. For example, out of 110 listed funders on the JULIET 2 service, 32 have data policies of some form. The topic has been discussed by national and international organisations, for example, ICSU (the International Council for Science), the OECD (Organisation for Economic Co-operation and Development) and the UK’s Royal Society. The public statements that emerge from these scientific bodies call for both research transparency and freely available access to research data created with public funding for possible reuse. The rights associated with the sharing of data and the environment in which it can be done is also of interest to publishers. This interest can be attributed to two motivating factors: to support the academic function of data such as the corroboration of research findings and the facilitation of the re-use of data; and to respond to a strategic, commercial development, for instance, an engagement with the rights, process and environment of data sharing. Currently some publishers are introducing contractual policies on the archiving and sharing of data in addition to policies governing the deposit and sharing of research articles through repositories. The issue of policies on sharing set out by academic journals has been raised by scientific organisations, such as the US National Academy of Sciences, which urges journals to make clear statements of their sharing policies. On the other hand, the publishing community whilst broadly supporting the principle of open and accessible research data expresses concerns over the intellectual property implications of archiving shared data.

Keywords. Journal Data Policies, Research Data Policies

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

¹ Corresponding Author: Marianne Bamkin, University of Nottingham, Centre for Research Communications; E-mail: Marianne.bamkin@nottingham.ac.uk.
Opening up Digital Publications –
Suggestions for Making Humanities Data Available for Research

Henriette ROUED-CUNLIFFE

Bavarian Academy of Sciences and Humanities, University of Munich

Abstract. Currently the trend in digital publishing of Humanities data seems to be moving towards openness and interoperability. In this abstract I will examine to what extent and in what way current digital publications are open and accessible. My hypothesis is that while many digital publications are currently made available online and can be searched and viewed by the general public, very few are available to researchers in a meaningful way. By meaningful I mean that external researchers can search and export data for reuse and are possibly even encouraged to define their own search criteria. I believe that this is the true essence of data sharing. Following this, I will propose one approach, using XML and Web Services, to creating a digital publication of Humanities data that would be open to the research community in a meaningful way, as defined above.

Keywords. External Researchers, XML, Web Services

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

1 Corresponding Author: Henriette Roued-Cunliffe, Ludwig-Maximilians Universität, Munich, Germany; E-mail: henriette@roued.com.
Repositories Recreated – Working Towards Improved Interoperability and Integration by a Co-Operative Approach in Sweden

Stefan ANDERSSON¹,² and Aina SVENSSON²

¹Electronic Publishing Centre, Uppsala University Library

Abstract. Recently the technological and organizational infrastructures of institutional repositories have been questioned. For example the British so-called Finch report from last summer argued that further development, as well as higher standards of accessibility of repositories, are needed in order to make them better integrated and interoperable to ultimately bring greater use by both authors and readers. Not only the technical frameworks and presumably low usage levels are criticized but also the lack of “clear policies on such matters as the content they will accept, the uses to which it may be put, and the role that they will play in preservation”. The report concludes that: “In practice patterns of deposit are patchy”.

As in the UK, today, all universities and university colleges in Sweden, except a couple of very small and specialized ones, do have an institutional repository. A majority (around 80%) are working together on a co-operative basis within the DiVA Publishing System with the Electronic Publishing Centre at Uppsala University Library acting as the technical and organizational hub. Because the system is jointly funded, and the members contribute according to their size, it has been possible even for smaller institutions with limited resources to run a repository with exactly the same functionalities as the biggest universities.

In this presentation we want to demonstrate the ever-increasing importance of institutional repositories in Sweden. Starting more than a decade ago the DiVA Consortium has, for some time, been addressing the problems now raised by the Finch report in a number of areas.

Keywords. Research Information, Long-term Preservation, DiVA, Sweden

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

¹ Corresponding Author: Stefan Andersson, Uppsala University Library, Sweden; E-mail: Stefan.andersson@ub.uu.se.
Towards an e-Publishing Library Service in Humanities and Social Sciences: A Feasibility Study

Rosa DI CESARE\textsuperscript{a,1}, Silvia GIANNINI\textsuperscript{a} and Daniela LUZI\textsuperscript{b}

\textsuperscript{a}National Research Council, Institute for Research on Population and Social Policies, Rome, Italy
\textsuperscript{b}National Research Council, Institute of Information Science and Technology, Pisa, Italy

Abstract. Introduction. Many surveys have been promoted and/or supported by SPARC (Scholarly Publishing and Academic Resources Coalition) and ARL (American Research Libraries) to gain understanding of the initiatives and projects undertaken by universities and/or research institutions in the development of value-added services to manage in-house publications. These surveys considered the advantages of using open source software to manage digital contents and pointed out critical issues in the development of these services, stressing the necessity of integrating them with other digital repositories. Case studies of successful strategies are also reported, highlighting the research context, type of products and/or collections to be managed with e-publishing services.

Aim. This paper intends to describe a methodology used to analyze the editorial production of CNR Institutes belonging to the Department of Humanities and Social Sciences. This analysis is considered a pre-requisite to design a feasibility study with the aim of developing an e-publishing service tailored to HSS characteristics. To reach this aim the paper describes in particular the characteristics of editorial products defining a set of quality criteria for current production. The result of this analysis can provide insight into the identification of weak and strong points that have to be addressed when developing a new and sustainable e-publishing service.

Survey design. To gain insight into the characteristics of editorial products we identified a set of variables that express stability (start date, number of years and frequency of publication); editorial quality (presence of standardized bibliographic elements and codes; attribution of Copyright/Creative commons; Peer-review process) and visibility (indexed in national/international catalogues and/or archives, access modes). For the purpose of our analysis the results of the survey are described distinguishing between the editorial products entirely managed in-house and those that are published and/or distributed by commercial publishers. Moreover, results are also reported by type of editorial products (Monograph series, journals and e-journals and report series) considering that each type of product has specific modes of publishing and editorial process.

Results. CNR Institutes in HSS produce different types of editorial products in a stable way and with continuity over time. There is a consistent number of series that have published for more than 20 years and editorial activity is keeping pace with new-born products that also include e-Journals. No major differences emerged in the editorial quality of in-house and external products, especially if we consider formal editorial aspects. The selection of content depends on the type of

\textsuperscript{1} Corresponding Author: Rosa Di Cesare, National Research Council, Institute for Research on Population and Social Policies, Rome, Italy; E-mail: r.dicesare@irpps.cnr.it.
products, while content evaluation of in-house publishing is not so diffused as well as the attribution of copyright/creative commons. The introduction of an e-publishing service could support more efficiently the peer-review process and also improve visibility thanks to their additional services embedded in their platform that support all the activities connected with the content exposure and retrieval in indexing and abstracting services. This is particularly important in HSS.

In the future we intend to further analyze the organization context where editorial activities are managed carrying out a questionnaire-based survey to explore the role of libraries and/or other stakeholders involved in this process as well as researchers’ needs when publishing their results.

**Keywords.** Journal Data Policies, Research Data Policies

Full text available at: http://elpub.scix.net/cgi-bin/works/Home
Making Social Interactions Accessible in Online Social Networks

Fredrik ERLANDSSON\textsuperscript{a,1}, Roosbeh NIA\textsuperscript{b}, Henric JOHNSON\textsuperscript{a} and Felix WU\textsuperscript{b}

\textsuperscript{a}Blekinge Institute of Technology, Karlskrona, Sweden
\textsuperscript{b}University of California, Davis, USA

\textbf{Abstract.} Recently, online social networks, OSNs, have gained significant popularity and are among the most popular ways to use the Internet. Additionally, researchers have become more interested in using the social interaction networks, SINs\textsuperscript{[1]}, in order to further enhance and personalize their services\textsuperscript{[2]}. OSNs are also redefining roles within the publishing industry, allowing publishers and authors to reach and engage with readers directly\textsuperscript{[3]}. However, SINs are not very easily available as of today through the current APIs provided by most OSNs. Such applications would therefore spend tremendous amount of time trying to gather the required SINs for their services. Therefore, our research problem is how we can design a system that makes social interactions in OSNs accessible. This also refers to the problem of how to crawl OSNs in a structured way, which is the focus of this short paper.

\textbf{Keywords.} Data Mining, Crawling, Social Interactions

Full text available at: http://elpub.scix.net/cgi-bin/works/Home

\textsuperscript{1} Corresponding Author: Fredrik Erlandsson, Blekinge Institute of Technology, 371 79 Karlskrona; E-mail: Fredrik.erlandsson@bth.se.
This page intentionally left blank
### Subject Index

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D interaction</td>
<td>13</td>
</tr>
<tr>
<td>3D visualization and navigation</td>
<td>13</td>
</tr>
<tr>
<td>7th framework programme</td>
<td>77</td>
</tr>
<tr>
<td>adaptive hypertext</td>
<td>3</td>
</tr>
<tr>
<td>algorithms</td>
<td>3</td>
</tr>
<tr>
<td>article-level metrics</td>
<td>77</td>
</tr>
<tr>
<td>clustering</td>
<td>13</td>
</tr>
<tr>
<td>conference management system</td>
<td>45</td>
</tr>
<tr>
<td>controled vocabularies</td>
<td>89</td>
</tr>
<tr>
<td>copyright</td>
<td>23</td>
</tr>
<tr>
<td>cover sheets</td>
<td>23</td>
</tr>
<tr>
<td>crawling</td>
<td>97</td>
</tr>
<tr>
<td>data mining</td>
<td>97</td>
</tr>
<tr>
<td>DCAP</td>
<td>33</td>
</tr>
<tr>
<td>deduplication</td>
<td>57</td>
</tr>
<tr>
<td>design</td>
<td>3</td>
</tr>
<tr>
<td>design science research</td>
<td>33</td>
</tr>
<tr>
<td>development of DCAP</td>
<td>33</td>
</tr>
<tr>
<td>digital humanities</td>
<td>67</td>
</tr>
<tr>
<td>digital libraries</td>
<td>67</td>
</tr>
<tr>
<td>digital repositories</td>
<td>91</td>
</tr>
<tr>
<td>DiVA</td>
<td>94</td>
</tr>
<tr>
<td>Dublin core</td>
<td>57</td>
</tr>
<tr>
<td>Dublin core application profile</td>
<td>33</td>
</tr>
<tr>
<td>experimentation</td>
<td>3</td>
</tr>
<tr>
<td>external researchers</td>
<td>93</td>
</tr>
<tr>
<td>force-directed placement</td>
<td>13</td>
</tr>
<tr>
<td>full text</td>
<td>67</td>
</tr>
<tr>
<td>geographic distribution</td>
<td>91</td>
</tr>
<tr>
<td>H.3.3 information search and retrieval/information filtering</td>
<td>3</td>
</tr>
<tr>
<td>H.5.1 multimedia information systems/evaluation/methodology</td>
<td>3</td>
</tr>
<tr>
<td>H.5.4 hypertext/hypermedia/theory</td>
<td>3</td>
</tr>
<tr>
<td>I.2.0 general/cognitive simulation</td>
<td>3</td>
</tr>
<tr>
<td>I.2.6 learning/connection simulation and neural nets</td>
<td>3</td>
</tr>
<tr>
<td>information retrieval</td>
<td>13</td>
</tr>
<tr>
<td>institutional repositories</td>
<td>23, 90</td>
</tr>
<tr>
<td>internet visibility</td>
<td>90</td>
</tr>
<tr>
<td>journal data policies</td>
<td>92, 95</td>
</tr>
<tr>
<td>k-means</td>
<td>13</td>
</tr>
<tr>
<td>linked open data</td>
<td>45, 89</td>
</tr>
<tr>
<td>long-term preservation</td>
<td>94</td>
</tr>
<tr>
<td>Me4DCAP</td>
<td>33</td>
</tr>
<tr>
<td>metadata</td>
<td>33, 57</td>
</tr>
<tr>
<td>method</td>
<td>33</td>
</tr>
<tr>
<td>OAI-PMH</td>
<td>57</td>
</tr>
<tr>
<td>open access</td>
<td>57, 77</td>
</tr>
<tr>
<td>open bibliographic data</td>
<td>45</td>
</tr>
<tr>
<td>OpenAIRE</td>
<td>77</td>
</tr>
<tr>
<td>PLOS</td>
<td>77</td>
</tr>
<tr>
<td>portals</td>
<td>67</td>
</tr>
<tr>
<td>R</td>
<td>77</td>
</tr>
<tr>
<td>record linkage</td>
<td>57</td>
</tr>
<tr>
<td>relative citation impact</td>
<td>90</td>
</tr>
<tr>
<td>research data policies</td>
<td>92, 95</td>
</tr>
<tr>
<td>research evaluation</td>
<td>77</td>
</tr>
<tr>
<td>research information</td>
<td>94</td>
</tr>
<tr>
<td>research infrastructures</td>
<td>67</td>
</tr>
<tr>
<td>retrieval infrastructures</td>
<td>3</td>
</tr>
<tr>
<td>semantic web</td>
<td>33</td>
</tr>
<tr>
<td>SKOS</td>
<td>89</td>
</tr>
<tr>
<td>social interactions</td>
<td>97</td>
</tr>
<tr>
<td>spreading activation</td>
<td>3</td>
</tr>
<tr>
<td>statistical computing</td>
<td>77</td>
</tr>
<tr>
<td>survey</td>
<td>23</td>
</tr>
<tr>
<td>Sweden</td>
<td>94</td>
</tr>
<tr>
<td>text analysis</td>
<td>23</td>
</tr>
<tr>
<td>theory</td>
<td>3</td>
</tr>
<tr>
<td>web services</td>
<td>93</td>
</tr>
<tr>
<td>XML</td>
<td>93</td>
</tr>
</tbody>
</table>
This page intentionally left blank
## Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page 1</th>
<th>Page 2</th>
<th>Page 3</th>
<th>Page 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abente, M.J.</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amado Alves, M.</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andersson, S.</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bamkin, M.</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baptista, A.A.</td>
<td>33, 89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berndt, R.</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caldera, C.</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creel, S.</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curado Malta, M.</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de Antonio, A.</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Di Cesare, R.</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erlandsson, F.</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fellner, D.W.</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fenner, M.</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmino, H.N.</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freire, N.</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giannini, S.</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hubbard, B.</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hussein, A.</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jahn, N.</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johnson, H.</td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klepec, D.</td>
<td></td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lavesson, N.</td>
<td></td>
<td></td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>Linde, P.</td>
<td></td>
<td></td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Luzi, D.</td>
<td></td>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>Moral, C.</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>Nia, R.</td>
<td></td>
<td></td>
<td></td>
<td>97</td>
</tr>
<tr>
<td>Olsbo, P.</td>
<td></td>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Polydoratou, P.</td>
<td></td>
<td></td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Roued-Cunliffe, H.</td>
<td></td>
<td></td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>Schirrwagen, J.</td>
<td></td>
<td></td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>Smith, J.</td>
<td></td>
<td></td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>Sturges, P.</td>
<td></td>
<td></td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>Svensson, A.</td>
<td></td>
<td></td>
<td></td>
<td>94</td>
</tr>
<tr>
<td>Taylor, S.</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Tonkin, E.L.</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Tourte, G.J.L.</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Welsh, T.S.</td>
<td></td>
<td></td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>Wendykier, P.</td>
<td></td>
<td></td>
<td></td>
<td>57</td>
</tr>
<tr>
<td>Wu, F.</td>
<td></td>
<td></td>
<td></td>
<td>97</td>
</tr>
</tbody>
</table>
This page intentionally left blank
This page intentionally left blank
This page intentionally left blank