

An International Journal on Grey Literature



Autumn 2015 - TGJ Volume 11, Number 3

**‘TOPICAL AND TECHNICAL
ADVANCES IN GREY LITERATURE’**

GreyNet

The Grey Journal

An International Journal on Grey Literature

COLOPHON

Journal Editor:

Dr. Dominic J. Farace
Grey Literature Network Service
GreyNet, The Netherlands
journal@grey.net.org

Associate Editors:

Julia Gelfand
University of California, Irvine
UCI, United States

Dr. Debbie L. Rabina
School of Information and Library Science
Pratt Institute, United States

Dr. Joachim Schöpfel
Université Charles de Gaulle Lille 3
France

Gretta E. Siegel
Portland State University
PSU, United States

Kate Wittenberg
Project Director, Client and Partnership
Development at Ithaka, United States

Technical Editor:

Jerry Frantzen, TextRelease

CIP

The Grey Journal (TGJ): An international journal on grey literature / D.J. Farace (Journal Editor); J. Frantzen (Technical Editor) ; GreyNet, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 11, Number 3, 2015. - DANS-KNAW, CVTISR, EBSCO, FEDLINK-Library of Congress, Inist-CNRS, ISTI-CNR, KISTI, NIS-IAEA, NTK, and NYAM are Corporate Authors and Associate Members of GreyNet International. This serial publication appears three times a year - in spring, summer, and autumn. Each issue is thematic and deals with one or more related topics in the field of grey literature. The Grey Journal appears both in print and electronic formats.
ISSN 1574-1796 (Print)
ISSN 1574-180X (E-Print/PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:

€240 institutional, including postage & handling

Contact Address:

Reprint Service, Back Issues, Document Delivery,
Advertising, Inserts, Subscriptions:

TextRelease
Javastraat 194-HS
1095 CP Amsterdam
Netherlands
T/F +31 (0) 20 331.2420
info@textrelease.com
<http://www.textrelease.com/gjpublications.html>

About TGJ

The Grey Journal is a flagship journal for the international grey literature community. It crosses continents, disciplines, and sectors both public and private. The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.



The Grey Journal is geared to Colleges and Schools of Library and Information Studies, as well as, information professionals, who produce, publish, process, manage, disseminate, and use grey literature e.g. researchers, editors, librarians, documentalists, archivists, journalists, intermediaries, etc.

About GreyNet

The Grey Literature Network Services was established in order to facilitate dialog, research, and communication between persons and organizations in the field of grey literature. GreyNet further seeks to identify and distribute information on and about grey literature in networked environments. Its main activities include the International Conference Series on Grey Literature, the creation and maintenance of web-based resources, a moderated Listserv, and The Grey Journal. GreyNet is also engaged in the development of distance learning courses for graduate and post-graduate students, as well as workshops and seminars for practitioners.

Full-Text License Agreement

In 2004, TextRelease entered into an electronic licensing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The full text of articles in The Grey Journal (TGJ) can be found in *Library, Information Science & Technology Abstracts* (LISTA) full-text database.

© 2015 TextRelease

Copyright, all rights reserved. No part of this publication may be reproduced, stored in or introduced into a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without prior permission of the publisher.

Contents

'TOPICAL AND TECHNICAL ADVANCES IN GREY LITERATURE'

An Attempt to Nuance the Understanding of Professional Reports in Archaeology	137
<i>Lisa Börjesson (Sweden)</i>	
A Reassessment of Archaeological Grey Literature: semantics and paradoxes	145
<i>Tim N.L. Evans (United Kingdom)</i>	
Recent Italian regulations about the open-access availability of publicly funded research publications, and the documentation landscape in astrophysics	167
<i>Monica Marra (Italy)</i>	
Marine Planning and Service Platform; an advanced research engine for grey literature in marine science	171
<i>Sara Goggi, Monica Monachini, Francesca Frontini, Roberto Bartolini, Gabriella Pardelli, Maurizio De Mattei, Franco Bustaffa, and Giuseppe Manzella (Italy)</i>	
Enhancement of the functions of the Japan Atomic Energy Agency Library's Fukushima Nuclear Accident Archive using a novel data flagging system that improves the utilization of numerical data on the Internet	179
<i>Kiyoshi Ikeda, Mayuki Gonda, Shun Nagaya, Misa Hayakawa, Yukinobu Mineo, Katsuhiko Kuni, Minoru Yonezawa, and Keizo Itabashi (Japan)</i>	
The Use of Web-Scraping Software in Searching for Grey Literature	186
<i>Neal R. Haddaway (Sweden)</i>	

Colophon.....	134
Editor's Note.....	136
On the Newsfront	
GL17 Call for Posters : Seventeenth International Conference on Grey Literature.....	192
GreyForum 4.1 : Grey Literature and Digital Preservation.....	193
Advertisements	
EBSCO LISTA Full-Text.....	166
FEDLINK, Federal Library and Information Network - Library of Congress.....	191
Author information.....	194
Notes for Contributors.....	195

Leveraging Grey Literature – A Case Study

Sustained improvements and enhancements in the production and access to grey literature are required. Leveraging grey literature not only seeks to capitalize on its value for science and the public good, but also anticipates a return on investments in some measure. In an attempt to assess GreyNet's return on investment in grey literature a study was carried out to identify and review efforts made to leverage its information products and services.

The method of approach is a cumulative case study from 2003 onwards of a select number of information products and services provided by GreyNet. Each were introduced to serve the interests of the grey literature community and their production and supply remain sustained. The first step in the study provides a brief description of nine selected resources and their intended purpose in serving the grey literature community. Available statistics and data applicable to each of the resources is also incorporated in this step. The second step was the design and implementation of an online survey among GreyNet's stakeholders in an effort to determine their satisfaction. And, in the third and final step, an analysis is carried out of the intended purposes and observed uses of the resources in step one with the results of the survey conducted among the stakeholders in step two.

The results from this study should establish an acceptable level of disparity between the observed uses and the expectations of stakeholders. The outcome should better enable GreyNet to leverage its grey literature resources – thus ensuring its positional advantage and enabling it to function more effectively.

In an effort to determine its potential application for other organizations both within and outside the grey literature community, preliminary results from the research project were presented as a training module to a group of federal librarians and information professionals at the Library of Congress. The final results of the project will be presented at the Seventeenth International Conference on Grey Literature this December in Amsterdam.

Dominic Farace
journal@greynet.org

An Attempt to Nuance the Understanding of Professional Reports in Archaeology*

Lisa Börjesson (Sweden)

Abstract

Professional (i.e. extra-academic) contract archaeology is an internationally widespread practice contributing significantly to the archaeology literature. However, professional knowledge production in archaeology, and most notably the professional report genre, is at times described as problematic. The problem descriptions are ambiguous and can be grouped under at least three different topics: concerns for content quality and practical accessibility, concerns for the comparably low degree of analytical and theoretical synthesizing in reports and concerns for lack of mutual knowledge transfer between academic archaeologists and professional archaeologists. Technical issues of access are to an increasing extent being solved. Format standardizations are also developing. Hence the report genre becomes more accessible, and the content more readable and informative. Yet articulations of attitudes toward the genre in archaeology text books and journal articles remain focused on the genre's problems. The aim of my ongoing dissertation research is to nuance the understanding of the professional report genre in archaeology. I do so by analyzing factors shaping reporting as it takes place in the intersection between academic norms, professional values and market logics. I argue an improved genre understanding is crucial to diminish cultural issues of access to the report literature, and also as a basis for development of reporting practices. In the dissertation research I analyze (1) perceptions about the report genre in archaeology literature, (2) information policy regulating reporting in archaeology, (3) how report writers and county board professionals interpret the reporting and report auditing work tasks and (4) the frames of reference report writers bring into reporting. The aim of this paper is to explicate the research design consisting of four sub-studies, to briefly report on findings from study no. 4, and to discuss preliminary, partial results from study no. 2.

Keywords: Documentation; Report; Professional report; Grey literature; Archaeology; Cultural heritage; Professional archaeology; Contract archaeology; Public government; Public administration; Cultural heritage management; Information policy; Science and technology studies; Scholarly communication; Sociology of professions; Mixed methods; Idea analysis; Policy analysis; Qualitative document analysis; Bibliometrics; Focus groups interviews

1. Introduction

The rise and development of professional (i.e. non-academic) archaeology during the twentieth century has led to a situation where professional archaeology constitute a substantial part of all archaeological activities (Aitchison 2010; Kristiansen 1998; cf. Ambrosiani 2012). Professional contract archaeology has increasingly contributed to archaeology literature, primarily by producing reports from surveys and excavations undertaken prior to land development (Aitchison 2010; Ersgård 2006). Digital means for report production, dissemination, and archiving have made reports increasingly visible and accessible to users.

The professional report genre's role as a contribution to archaeological knowledge is subject to scrutiny (e.g. by Bahn 2012; Hodder 1989; cf. Lucas 2001). Discussions about the role of reports can be followed for example in two special issues of research journals around 2010: in *Archaeologies* from the viewpoint of archaeology and in *The Grey Journal* from the viewpoint of information science (both edited by the archaeologist Seymour 2010a; 2009a). These special issues cover topics like professional review (e. g. Harlan 2009), archiving and access (e. g. Kansa et al. 2010; Aitchison 2009), but also expound the relationship between professional archaeology and academic archaeology and the transfer of knowledge between the two (e. g. Aitchison 2010; Roth 2010; Seymour 2009b). Taken together the descriptions allude to a compound of issues where the problem focus at times lies within the genre, and at other times lies with attitudes towards the genre.

My ongoing dissertation research takes these ambiguous problem descriptions as a departure point and seeks to nuance the description of the archaeology report genre. The nuancing takes on the perspective of contract archaeology as a professional practice, embedded in heritage policy, administrative requirements, market logics, organizational conditions and professional

* First published in the GL16 Conference Proceedings, March 2015.

ideals. The aim is, through an exploration of contextual factors shaping reporting to provide more informed grounds for readings and evaluations of reports. This re-contextualisation will serve as a basis to diminish some of the cultural issues of access to the genre, and from which to manage development of the report genre from policy and organizational levels. The overarching dissertation research questions are:

- I. Which ideas about the report genre are articulated in discussions of 'the grey literature problem' in archaeology?
- II. How are professional archaeology reports shaped by contextual factors such as heritage policy, market logics, administrative requirements, organizational conditions and professional ideals?
- III. Can insights from studies of reporting contexts be used to address some of the issues articulated in the discussions of 'the grey literature problem'?

In this paper I lay out my mixed methods dissertation research design and report on findings from two of the four sub-studies. The paper begins by framing archaeological reporting as a genre of scholarly communication. The first section also explains my operationalization of the concept 'configuring factors' of documentation. The four sub-studies making up the dissertation research are described subsequently. The paper concludes by a discussion of the methodological approach and related ethical considerations on researching a professional discipline and its activities.

2. A genre and cultural issues of access

Throughout this paper I conceptualize archaeological reports as a genre. The use of the genre concept brings with it a grouping of individual documents. A library and information science (LIS) genre concept can also lead us to focus on the relations between documents, communication, activities and social organization (Andersen 2008, 339). Genres are assumed to not only be created in social settings, but to also create the settings by influencing structures for communication and actions (2008, 350). This view on the genre concept implies that reports shape archaeology, which motivates the dissertation research from an archaeology perspective.

If we begin with the understanding that the archaeology report genre is a boundary genre, dependent on both government heritage management and archaeology, we can assume the report genre is conditioned by multiple factors and therefore hard to describe once and for all (cf. Börjesson under review; Huvila 2011). Heritage management logics and archaeology develop both independently and interrelated and each make claims as to how construct archaeology documentation should be done. Furthermore the market organization of professional archaeology has brought about a multitude of organizational forms among the actors at the market, affecting the organizational contexts in which documentation is carried out (cf. Aitchison 2010).

Following the assumption that the report genre as a boundary genre is conditioned by multiple factors, it is likely that the multiple problems related to archaeological reporting not will be solved by any one measure. The genre's problems have been thoroughly problematized and argumentation pro and con the genre has been spelled out (Seymour 2010a; 2009a). The problems described are such as inadequate archiving, instable archives, survey data being disconnected from reports, perceptions about lack of peer-review, lack of interest in and respect of report content from academic archaeologists, and language barriers. I (tentatively) argue several of these problems arise when reports are read and evaluated in relation to academic ideals as a standard. One way forward could be, based on the assumptions that reporting is configured by multiple factors on the boundary between academic, governmental-professional and market logics, to seek further understanding of the non-academic factors shaping the reporting genre in professional archaeology. This approach would offer a model for understanding why reports become what they are and, by doing so, provide report readers with a firmer ground from which to read and interpret report contents. Furthermore a model based on factors shaping professional archaeology reporting will provide a basis for evaluations and management of reporting practices.

The genre concept emphasizes common traits across documents. In order to nuance the understanding of the report genre I hence complement the genre perspective with the concept 'configuring factors' (of documentation) borrowed from Bernd Frohmann. Frohmann proposes a four-fold perspective to cover the "the configuring factors" of documents: the materiality of

documents, their histories, the institutions in which they are embedded, and the social discipline shaping practices with them (Frohmann 2004, 405). This dissertation covers the last three mainly social factors, while leaving out the materiality. To explore configuring factors serves as a way to study factors shaping the genre. Variations within the genre are particularly interesting for a nuanced understanding of the genre. In my use of the term 'configuring factors' I therefore make the assumption that factors shaping a genre have unequal impacts different parts of the genre, resulting in variations of interest for overall understanding of the genre.

In a larger perspective this study can be placed in the field Science and Technology Studies (STS), particularly in the sub-field focusing on scientific (or 'scholarly') communication (Sismondo 2010). Professional archaeology reports are not simply scientific – they balance on the boundary between academia and government. Nevertheless information policy regulating the report genre expresses aspirations for the genre to be scholarly and to contribute to scientific knowledge (Riksanstikvarieämbetet 2012). The boundary position is sometimes framed by the term 'grey literature' (Darvill 2008). Grey literature is a cross-disciplinary term used to describe literature produced on all levels of government, academics, business and industry where publishing is not the primary activity of the producing body (Farace and Schöpfel 2009).

3. Do we have a problem? Articulations of 'the grey literature problem' in archaeology literature (Study no. 1)

Articulations of the 'grey literature problem' in archaeology are ambiguous. They include descriptions of technical access challenges (e.g. reports are stored in regional depositories, Silva 2010), content usability issues (e.g. content is not validated by data, Roth 2010), and readers' confused expectations on the genre (e.g. false perceptions about lack of peer-review, Harlan 2010). These articulations point out deficiencies both within the genre and among its readers. A number of the voices also suggest ways to improve reporting and report use. Some of the articulations specify which national or other limitation of context they cover, yet other speaks about the problems in general terms.

This article analyzes general (mainly Anglo-American) articulations of the problems with the report genre in archaeology and how these problem articulations are reflected in Swedish archaeology literature. The aim is to explore nuances in the problem description and related 'solutions'. The guiding questions are: How are the challenges with grey literature described in the articles in the two special issues on professional archaeology documentation *Archaeologies* 2010 and *The Grey Journal* 2009? Are these articulations reflected in contemporary texts in Swedish? If so, what do the Swedish descriptions pick up and what do they leave out? The underlying assumption is that the problem descriptions not only describe a problem, but also foreshadow terms and forms for describing a phenomenon that is not necessarily in itself a problem. Hence the study does not seek to reinforce the problem, but to illuminate a cluster of ideas about the report genre in a reflective manner.

STS has a tradition of addressing problems or controversies. Controversies can be approached as a way to analyze a discipline at a point in history where conflicting perspectives intersect, are articulated and thus form rhetoric about phenomena, facts or artifacts. Another term used as a metaphor for this process is 'black boxing'. The STS study of controversies focuses at stages when a black box is open, when a phenomenon is about to be defined, or a previous dominating paradigm is being contested. As the controversy is settled the black box closes and the 'winning' argument become neutralized as 'knowledge'. The STS norm is to apply a symmetrical approach where all sides in conflict are analyzed as reasonable from their own perspective (Sismondo 2010). Important to note is that the ambiguous descriptions of the grey literature problem in archaeology does not necessarily qualify as a controversy in the sense that we see distinct opposite sides openly arguing for their own standpoint and against each other's. The grey literature problem in archaeology has a less outspoken character, but yet receives attention from a number of researchers and professional archaeologists from different angles (Seymour 2010a; Seymour 2009a).

The material will consist of articulations made by article authors in the two special issues of *Archaeologies* 2010 and *The Grey Journal* 2009, both on the topic grey literature in archaeology. The articulations made by the article authors will be compared to ideas in a selection of Swedish texts concerning the topic professional knowledge production and related challenges. The analysis method will be a qualitative document analysis (QDA) specifically focused on dimensions in the problem descriptions (Bergström and Boréus 2012; Plano Clark et al. 2010). The focus on

dimensions in the problem descriptions is a way to encompass both the 'problems' and the 'solutions' written out or alluded to by the authors in the material.

4. Negotiating documentation in professional practices - The case of archaeology reporting (Study no. 2)

Information policy in heritage legislation and guidelines is the primary means by which the government can handle the report genre challenges. Secondary means for the government to use can be supporting academic heritage education and stimulating professional education among archaeologists and county administrative board professionals. Information policy comprises decisions, guidelines, regulations and laws directly involving information creation, processing, flows and use (Braman 2011). Information policy regarding reporting articulates ideas and norms on what reports should be like, hence how reports shall respond to the needs of all of the stakeholders whose voices have been heard in the policymaking process.

Heritage policy is comprised of both hard law on for example, heritage interests and property rights and on international antiquities trade, and soft law on how surveys should be documented and how documentation subsequently should be handled. The hard law portions of heritage legislation has assigned enforcement institutions such as the property and environment court in the case of property rights and the border police in the case of antiquities trade. Enforcement of soft law is based upon non-legally binding regulations and guidelines by government authorities and non-governmental organizations in a field. There is also a moral responsibility resting on archaeology and heritage management professionals, and on professional associations to act according to regulations.

Preliminary results from an analysis of information policy in Swedish cultural heritage policies show a low level of regulation of archaeology reporting in the heritage legislation. The county administrative boards are assigned the juridical responsibility to ensure adequate documentation and reporting, and also to set standards for what is being seen as 'adequate'. County administrative boards request reporting in public tenders, but also require archaeology contractors to establish documentation plans. The practical governance model hence distributes the responsibility between the county boards as clients and auditors, and the archaeological contractors as executors (Börjesson, Petersson, and Huvila work in progress). A further analysis will focus on the level of county administrative boards and archaeological contractors. The aim of the analysis is to explore how policies on reporting and professional responsibility play out in negotiations between public procurement tenders and archaeology contractors' documentation plans. QDA will be applied to structure the analysis and to pursue a discussion of how the ideas from the national level information policy play out in the interactions between county administrative boards and archaeological organizations in the case of documentation plans (Bergström and Boréus 2012; Nyqvist 2011; Cris Shore and Wright 2011; Plano Clark et al. 2010).

5. Professional interpretations of the reporting and report auditing work tasks – The case of professional archaeology (Study no. 3)

The third study explores archaeology professionals' interpretations of the reporting work task and government authority professionals' interpretations of the report auditing work task. While previous studies tracks the articulations of the grey literature problem in archaeology and information policy concerning reporting this study analyzes how reporting archaeologists and those overseeing the reporting practice formulate their view on their respective work task. The aim is to find out how professionals interpret reporting and report auditing as work tasks and responsibilities. The analysis will provide a basis for discussing professionals' points of view on reporting.

The material for this study will consist of results from previous studies of reporting (primarily government reports) and qualitative focus group interviews with county administrative board professionals and with professional archaeologists (at separate occasions). The two types of professionals will be given a similar set of points of discussion, but slanted toward their practices at 'opposite ends' of the reporting process. Results from the previous sub-studies may be incorporated into the material for the interviews.

The interview records will be analyzed in the light of theory of professions (Svensson and Evetts 2010). The key assumption upon which the choice of theory is based is that reporting is made in a primarily professional environment, outside academia and embedded in an organizational setting with more shaping forces like topical training, professional ideals and obligations toward

organization and colleagues etcetera, which interacts with the policy regulating reporting. Possible theoretical terms to operationalize in the interpretation are 'civic epistemologies', 'professional (or 'expert') knowledge' and 'information interests' (Mosse 2011; Jasanoff 2005; Sundin and Hedman 2005).

6. Grey literature – grey sources? Report writers' frames of reference (Study no. 4)

In this study I explore report writers' frames of reference in order to understand more about variations among report writers' perspectives (Börjesson under review).¹ Frames of references are studied by an analysis of source use patterns in bibliographical reports. The source use patterns are analyzed based on a detailed coding of cited sources in 97 Swedish field evaluation reports from 2013. The coding structure was developed iteratively to cover aspects represented in bibliographical lists. The variables catch source age, source type, source format, source language, organizational and spatial relation between source originator and report author.

Information source use in archaeology in general has been studied previously (Huvila 2014a; Huvila 2006). Use of specific types of sources has also received attention, specifically image use (Beaudoin 2014; Moser 2012). Isto Huvila's study target variations in source use among archaeologists related to work roles. My study focuses specifically on source use in professional reports and explores variations in source use across the variables through a correspondence analysis. The bibliometric approach provides means to analyze distributions within the dataset, to test connections between parts of data and to reveal latent patterns (Denscombe 2009, 327; Tague-Sutcliffe 1992, 1).

The study begins with the assumption that as archaeological reports primarily are written in extra-academic, professional settings, it is relevant to seek to understand report writers' frames of references in relation to professional standards colored by organizational affiliations (Kristiansen 1998). Organizational types are therefore a key aspect in the analysis and interpretation, aided by sociology of professions influenced LIS research discussing information cultures and information interests shaped by occupational identities (Widén-Wulff 2010; Sundin and Hedman 2005).

The findings partly confirm previous studies of information source use, but also show some notable discrepancies. Reports are a key source for report writers in addition to professional literature and maps. Notwithstanding maps, non-codex like information sources are rare in the bibliographical lists. In previous studies both archaeological materials, aggregates of original materials and images have been described as information sources vital to archaeologists (Beaudoin, 2014; Huvila, in press a; cf. Moser, 2012). The complete lack of references to archaeological materials and the very rare references to images suggest that report writers either do not use these sources or that there are circumstances hindering archaeologists from representing these materials as sources. The national character of source choices shown in previous studies is largely confirmed. The source age distribution analysis shows a strong correlation between source type and source age. The importance of novelty is in this case related to source type and also possibly to additional contextual issues as accessibility.

References to databases and webpages (not including map databases) are about as common as references to academic literature, like the results in Huvila's study (2014a). Two notable characteristics of the use of digital sources are the relatively few digital sources, and the very rare occurrences of references to sources other than codex-like sources. Digitization and born-digital information in other forms such as digital dataset, objects in digital collections, 3D models, virtual reality visualizations seem to be rarely or never represented.

The correspondence analysis used as a way to analyze source use patterns reveal one administrative, one professional/academic pattern, and one map pattern. These three clusters show that the main division between source use patterns in the report genre is the division between the administrative source use pattern and the professional/academic source use pattern. This finding implies that report authors primarily relate to two spheres of previous knowledge production: an administrative sphere containing planning documents and reports more closely related to local cultural heritage management and a professional/academic sphere comprising more of the type of sources traditionally recognized as academic, but also a range of non-academic professional publications. These two patterns suggest a variation within the report

¹ The dataset will be made available along with the article upon acceptance for publication.

genre, where the dividing line goes between administrative and professional/academic source use patterns.

This study contributes by an examination of archaeological information use in a specific context, which has been requested in previous research (Huvila, in press a; Moser, 2012; Kansa and Kansa, 2011). The results can be employed to nuance both the understanding of professional archaeologists' information use, and to refine the understanding of the report genre from an information source use point of view. Awareness of variations between frames of reference among report writers could alter the perception of reports as "grey literature", in the sense being not academic (cf. Seymour 2010b).

7. Researching factors shaping scholarly communication in the intersection between academia, government and market – A brief discussion of methodology and research ethics

Above I make the argument that professional reporting in archaeology is influenced by a range of factors and that these factors, at least to some extent, are overlooked in evaluations of the report genre. Within the scope of the dissertation research I explore four of these factors; ideas in articulations of the genre's problems, policy regulating the genre, report writers' and county board administrators' interpretations of the reporting and report auditing work tasks, and report writers' frames of reference. Borrowing an anthropological metaphor, these four factors can be seen as 'sites' at the professional reporting field (cf. Wright 2011). From that point of view the dissertation research become a ethnography of a professional documentation practice with the aim to nuance the understanding of the genre produced within the practice.

As for ethnographic research in general the framing of the research problem and the choice of sites from which to study the problem is pivotal. In the case of my research the frame of the research problem is the report genre challenges as expressed by archaeologists. The description of these challenge articulations and how general articulations travel to and reflect in the Swedish situation will be further detailed by study 1. The choice of sites, apart from the first case study illuminating the challenge articulations, follows a stylized timeline in the professional archaeological process: from heritage policy through professionals' interpretation of the work task and finally implementation. Moreover the choice of sites cover a macro-micro dimension; heritage policy regulate all archaeologists, interpretations of the reporting and report auditing work tasks are influenced by the policy level but mainly take form in local social interactions and finally the report writers' frames of references reflect the perceptions of previous knowledge individuals bring with them into report writing.

The ethical considerations impacting this research circulate around the notion of professional practice. Professionals at work have reasons for the way they do their work which should be acknowledged in the analysis. However these reasons may not always be flattering, and may collide with strivings for a professional identity (cf. Mosse 2011). Ethnographic researchers must be prepared to balance the goal of ethnographic research with a respect for those studied. Furthermore the stylized timeline on which this dissertation research is based cannot be seen as a mirror of the practice, but more of as a model upon which the public sector is structured (from policy to professional interpretation and through to implementation), and thus from which public sector processes can be analyzed. This type of ethnographic multi-sited research will never explain individual events, but seeks to understand a phenomenon on a higher degree of abstraction. Ethnographic research traditionally attempts to let the studied community have an opportunity to review the research, through some type of public dissemination (Mosse 2011). It is less common that ethnographic research seeks to arrive at suggestions for changing the particular practice studied. This study should plan for a communication with the practice studied, and should in my opinion also consider if there are practical suggestions to make based on the thesis research regarding management of reporting practices.

References

- Aitchison, Kenneth. 2009. "Standards and Guidance in Archaeological Archiving: The Work of the Archaeological Archives Forum and the Institute for Archaeologists" *The Grey Journal* 5 (2): 67–71.
- . 2010. "Grey Literature, Academic Engagement, and Preservation by Understanding" *Archaeologies* 6 (2): 289–300. doi:10.1007/s11759-010-9145-5.
- Ambrosiani, Björn. 2012. "Swedish Contract Archaeology in the 1950s and 1960s" In *Histories of Archaeological Practices*, edited by Ola W. Jensen, 305–24. Stockholm: Statens Historiska Museum.
- Andersen, Jack. 2008. "The Concept of Genre in Information Studies" In *Annual Review of Information Science and Technology*, 339–67. Medford, NJ: Information Today.
- Bahn, Paul G. 2012. *Archaeology: A Very Short Introduction*. Oxford: Oxford University Press.
- Beaudoin, Joan E. 2014. "A Framework of Image Use among Archaeologists, Architects, Art Historians and Artists" *Journal of Documentation* 70 (1): 119–47.
- Bergström, Göran and Kristina Boréus. 2012. "Idé- och ideologianalys" In *Textens mening och makt*, edited by Göran Bergström and Kristina Boréus, 149–80. Lund: Studentlitteratur.
- Börjesson, Lisa. under review. "Grey Literature - Grey Sources? Nuancing the Perception of Professional Documentation: The Case of Archaeology"
- Börjesson, Lisa, Bodil Petersson and Isto Huvila. work in progress. "Information policy for sustainable digitization in archaeology? – A critical analysis"
- Braman, Sandra. 2011. "An Introduction to Information Policy." *Information Policy*, no. 1: 1–5.
- Darvill, Timothy. 2008. *The Concise Oxford Dictionary of Archaeology*. Oxford: Oxford University Press.
- Denscombe, Martyn. 2009. *Forskningshandboken*. Lund: Studentlitteratur.
- Ersgård, Lars. 2006. "Det uppdragsarkeologiska alternativet" In *Det förflutna är att räkna med*, edited by Lars Ersgård, 10–25. Stockholm: Riksantikvarieämbetet.
- Farace, Dominic J. and Joachim Schöpfel. 2009. "Grey Literature" In *Encyclopedia of Library and Information Sciences, Third Edition*, 2029–39. Oxford: Taylor & Francis.
- Frohmman, Bernd. 2004. "Documentation Redux: Prolegomenon to (Another) Philosophy of Information" *Library Trends* 52 (3): 387–407.
- Harlan, Mark E. 2009. "Archaeology's Grey Literature and Professional Review" *The Grey Journal* 5 (2): 83–88.
- . 2010. "Black and White Literature, Grey Scholars." *Archaeologies* 6 (2): 270–88. doi:10.1007/s11759-010-9143-7.
- Hodder, Ian. 1989. "Writing Archaeology: Site Reports in Context" *Antiquity* 63: 268–74.
- Huvila, Isto. 2006. *The Ecology of Information Work: A Case Study of Bridging Archaeological Work and Virtual Reality Based Knowledge Organisation*. Åbo: Åbo akademis förlag.
- . 2011. "The Politics of Boundary Objects: Hegemonic Interventions and the Making of a Document" *Journal of the American Society for Information Science and Technology* 62 (12): 2528–39.
- . 2014a. "Archaeologists and Their Information Sources" In *Perspectives to Archaeological Information in the Digital Society*, edited by Isto Huvila, Meddelanden Från Institutionen För ABM. Uppsala: Department of ALM, Uppsala University.
- . 2014b. "Introduction" In *Perspectives to Archaeological Information in the Digital Society*, edited by Isto Huvila, Meddelanden Från Institutionen För ABM. Uppsala: Department of ALM, Uppsala University.
- Jasanoff, Sheila. 2005. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton, N.J.: Princeton University Press.
- Kansa, Eric C. and Sarah Whitcher Kansa. 2011. "Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing" In *Archaeology 2.0*, edited by Eric C. Kansa, Sara Whitcher Kansa and Ethan Watarall, 57–92. <http://escholarship.org/uc/item/1r6137tb#page-73>.
- Kansa, Eric C. et. al. 2010. "Googling the Grey: Open Data, Web Services, and Semantics" *Archaeologies* 6 (2): 301–26. doi:10.1007/s11759-010-9146-4.
- Kristiansen, Kristian. 1998. "Gamla gränser och nya frontlinjer" In *Arkeologiska horisonter*, edited by Håkan Karlsson, 176–99. Stockholm: Symposion.
- Lucas, Gavin. 2001. *Critical Approaches to Fieldwork: Contemporary and Historical Archaeological Practice*. London: Routledge.
- Moser, Stephanie. 2012. "Archaeological Visualization - Early Artifact Illustration and the Birth of the Archaeological Image" In *Archaeological Theory Today*, edited by Ian Hodder, 292–322. Cambridge: Polity.
- Mosse, David. 2011. "Politics and Ethics: Ethnographies of Expert Knowledge and Professional Identities" In *Policy Worlds: Anthropology and the Analysis of Contemporary Power*, edited by Cris Shore, Susan Wright and Davide Però, 50–67. New York: Berghahn Books.
- Nyqvist, Anette. 2011. "Sweden's National Pension System as a Political Technology" In *Policy Worlds: Anthropology and the Analysis of Contemporary Power*, edited by Chris Shore, Susan Wright and Davide Però, 205–22. New York: Berghahn Books.
- Plano Clark, Vicki L. et. al. 2010. "Mixing Quantitative and Qualitative Approaches - An Introduction to Emergent Mixed Methods Research." In *Handbook of Emergent Methods*, edited by Sharlene Nagy Hesse-Biber and Patricia Levy, 363–87. New York/London: The Guilford Press.
- Riksantikvarieämbetet. 2012. *Uppdragsarkeologi (2 Kap, 10-13 §§). Vägledning för tillämpning av kulturminneslagen*. Stockholm: Riksantikvarieämbetet. <http://www.raa.se/wp-content/uploads/2012/06/V%C3%A4gledning-Uppdragsarkeologi-avsnitt-1-9.pdf>.
- Roth, Barbara J. 2010. "An Academic Perspective on Grey Literature" *Archaeologies* 6 (2): 337–45. doi:10.1007/s11759-010-9141-9.
- Seymour, Deni J. ed. 2009a. "The Grey Journal - Archaeology and Grey Literature" 5 (2).
- . 2009b. "Special Ethical and Scholarship Considerations for the Archaeological Grey Literature." *The Grey Journal* 5 (2): 89–96.
- ed. 2010a. "Archaeologies - Black-and-White Issues About the Archaeological Grey Literature," no. 2.

- .2010b. "In the Trenches Around the Ivory Tower: Introduction to Black-and-White Issues About the Grey Literature." *Archaeologies: Journal of the World Archaeological Congress* 6 (2): 226–32.
- Shore, Cris, and Susan Wright. 2011. "Conceptualising Policy: Technologies of Governance and the Politics of Visibility" In *Policy Worlds: Anthropology and the Analysis of Contemporary Power*, edited by Cris Shore, Susan Wright, and Davide Però, 1–26. New York: Berghahn Books.
- Silva, Rosiclér Theodoro da. 2010. "Some Aspects of 'Grey Literature' in Brazil" *Archaeologies* 6 (2): 327–36. doi:10.1007/s11759-010-9142-8.
- Sismondo, Sergio. 2010. *An Introduction to Science and Technology Studies*. Chichester, West Sussex, U.K: Wiley-Blackwell.
- Sundin, Olof, and Jenny Hedman. 2005. "Theory of Professions and Occupational Identities." *Information Today*, 293–97.
- Svensson, Lennart G. and Julia Evetts. 2010. *Sociology of Professions: Continental and Anglo-Saxon Traditions*. Göteborg: Daidalos.
- Tague-Sutcliffe, Jean. 1992. "An Introduction to Infometrics" *Information Processing & Management* 28 (1): 1–3.
- Widén-Wulff, Gunilla. 2010. "Organizational Culture." In *Encyclopedia of Library and Information Sciences, Third Edition*, 4073–78. Oxford: Taylor & Francis.
- Wright, Susan. 2011. "Studying Policy: Methods, Paradigms, Perspectives" In *Policy Worlds: Anthropology and the Analysis of Contemporary Power*, 27–31. New York: Berghahn Books.

A Reassessment of Archaeological Grey Literature: semantics and paradoxes*

Tim N.L. Evans (*United Kingdom*)

Abstract

This article takes a fresh view of unpublished archaeological reports, common in archaeological practice in England since the advent of PPG16. Although these reports are almost ubiquitously referred to as 'grey literature', they are but a facet of a larger corpus of publication and dissemination techniques used by the archaeological community. It is argued that the term 'grey' has become synonymous with a liminal status and inferior quality that is in fact contrary to the realities of the increased online publication of fieldwork reports. The article also considers the current upsurge in research projects and academic theses using fieldwork reports, and highlights the necessity that findings from research are fed back into the curatorial sector and baseline data to inform all aspects of archaeological work rather than isolated in published literature often unavailable to those outside of academia. Finally, the article argues that although the challenge of increased access – via the web – is being met, this is not consistent across the country, leading to lacunae in the information landscape. Furthermore, as the number of online reports grows into the tens of thousands there is a need for greater sophistication and archaeological context in the accompanying metadata, to aid classification and reuse.

1.1 Introduction

For a term with somewhat underwhelming connotations, grey literature is a common and sometimes controversial topic within archaeological discourse (see Seymour 2010a). In recent years it has become synonymous worldwide with the unpublished reports created through development-led fieldwork, and thus the challenges of integrating the results of such fieldwork into curatorial and research practices (Gibbs and Holley 2012; Louwe Kooijmans 2009; Silva 2010). However, perceptions are beginning to change as outputs that were once considered liminal become an integral part of the research paradigm; their value for use within a new wave of large-scale data analysis and synthesis evident in any country with a history of rescue or planning-led archaeological mitigation (Bradley et al 2015; Habu 2004; Smyth 2014). In addition with the simultaneous development of digital archives it is noticeable that the discipline is moving away from traditional and insular concerns of publication backlog and archive redundancy, towards the pivotal role of grey literature within a new frontier of cross-border and data-intensive research (Ariadne 2014; Kansa et al 2010; Oikarinen 2014; Vlachidis and Tudhope 2013). It seems that the wheel has turned, and that grey, to stretch the cliché, is the new black. Whilst the grey revolution gains momentum, this paper hopes to offer some fresh perspectives based on recent developments and trends in the reporting of intrusive investigations in England since the advent of PPG16 (see Fulford 2011). The 'English' experience encapsulates the essence of the grey literature story, wherein a corpus that was once perceived as a pariah of the discipline has begun to move into the academic spotlight (cf. Fulford and Holbrook 2011; Hardman 2010; Lock 2008). However, in the rush to embrace this shadowy corpus of data there has arguably been a degree of naivety as to what is really meant, perceived, and understood about the term itself, and the realities of the dissemination of the results of contemporary fieldwork. Influenced by historic crises within archaeological publication, 'grey' has often become a throwaway yet somewhat restrictive label, as well as an epithet that reflects a lack of academic engagement with the outputs of the professional sector (c.f. Bradley 2006; Harlan 2010). Given this history, semantic controversy and increased online dissemination, it is pertinent to redefine what it means to be archaeologically grey, or if in fact, the term is still of use. In addition this paper aims to discuss some of the key issues and challenges for the use of reports literature within current archaeological practice. Contrary to popular perception, these are not the often subjective notions of the 'value' of content, but are the geographical and cultural biases in the numbers of reports in existence and an increasing disparity in online access. In addition, as period-based academic syntheses become common, there are also limitations in traditional methods of publication in ensuring that findings are communicated to the community at large. Accessing fieldwork reports is only the beginning, ensuring its reciprocal role within all aspects of archaeology is perhaps the greater challenge.

* First published in *Internet Archaeology*. 40 (2015). <http://dx.doi.org/10.11141/ia.40.6>

1.2 Data Sources

The following paper relies heavily on digital resources, primarily the Excavation Index, a guide to the archaeological excavations and interventions carried out in England since the earliest days of scientific archaeology (Archaeology Data Service 2011). The Index is part of the National Record of the Historic Environment (NRHE), and currently held online within the Archsearch database of the Archaeology Data Service (ADS), and via a web service through the Heritage Gateway. The dataset used in this study was obtained from the ADS as a set of delimited text files on 09/09/2014. The Index comprises digital information compiled by the Archaeological Investigations Project (AIP) and OASIS, as well as hard-copy information from journals, reports and monographs. As such, the database has historically been the most comprehensive national record for events in the country, and through its wide scope covers much of the shortfall encountered in the individual coverage of AIP and OASIS (Evans 2013; Holbrook and Morton 2008). The significance of the Index as an overarching source of events is perhaps understated in many works that deal with national syntheses and data intensive research. This is perhaps due to the fact that within the ADS and the Heritage Gateway it can only be searched on associated monument and geographic location, not the classification of the event itself or bibliographic sources. Although the Index can be obtained in its original form from the ADS, its current online incarnation, somewhat ironically, obscures its original purpose.

That is not to say however that the Index is definitive; previous study has shown that the Index contains lacunae, especially when compared to the event records of local Historic Environment Records (Evans 2013). In addition within the Index there is currently no record for events undertaken after 2011, reflecting its last update within the ADS (Hardman pers comm). Furthermore, there seems to be a shortfall in recorded investigations after the years 2007/8, presumably reflecting the ongoing updating of the database via sources such as OASIS and AIP. However despite these caveats, the Index still remains the only consistent *national* record for investigations. Although it may be assumed that local HER event records record more information, previous studies have shown that there are events and reports recorded only in AIP or OASIS, and thus the Excavation Index, but not the HER; as well as events that remain as an ongoing 'backlog' of information still to be entered into HERs (Brookes 2003; Evans 2013; Hardman and Evans 2010). Thus, in truth, no one source of events can ever be considered complete, the disparity between systems complicated not only by real gaps, but perhaps also by real or apparent discordance in classification of events (see Evans 2013).

Leaving aside issues of relative accuracy, the rationale in using the Excavation Index for a national study is simply one of practicality. There are currently 82 HERs in England, a mixture of HBSMR and bespoke systems (MacLean 2014), and although the majority are online via the Heritage Gateway, current functionality of this system does not extend to the searching of events or sources. The capacity to obtain and combine the individual datasets lies outside the immediate aims of this work, indeed a comparison of the holdings of local and national systems would make an interesting study for those interested in the epistemological and methodological connotations for the archaeological knowledge base (see Cooper and Green 2015). On this note, the current Historic England Heritage Information Access Strategy (HIAS) looks to increase interoperability between the NRHE and HERs, with enhanced capacity to transfer data to HERs and the enabling of these resources as the first port of call for researchers (see Boldrini et al 2015). It is hoped that future research into investigation and publication trends will be able to utilise a truly combined system of recording events and outputs, incorporating the strengths of local and national records, and thus allowing a more comprehensive analysis than currently available. The potential future developments aside, the Index will continue to play a role for researchers and will be used for this work, although the caveats highlighted above should be noted in any interpretation.

The final dataset used for this paper is the Library of Unpublished Fieldwork Reports of the ADS, the data cited in this paper was accessed 22/10/2014 as a download from the ADS. The dataset was restricted to only those records that were created through the OASIS system to reflect the current mainstream workflow within archaeology in England. It should be noted however that the library contains a large number of reports from other sources such as backlog scanning projects provided by curators and funders such as Lincolnshire County Council and the Highways Agency. However these have not been included within the dataset as they represent singular cases of specifically funded projects, not the everyday experience as exemplified by the use, or non-use, of the OASIS system. It should be noted at this point that use of the OASIS system by fieldworkers and curators in England has been uneven (Smith et al 2012). The data from the ADS

Library and OASIS is used here not as a 'complete' dataset, but conversely as an indicator of the disparities inherent in national patterns of recording.

2 Grey Literature revisited

2.1 A short history of grey and archaeology

The use of the term 'grey literature' within the archaeological vocabulary is an interesting historiographical study. It is not a discipline-specific term, but originates from the library and information science communities (see Falkingham 2005). The phrase itself has long held a fascination for all disciplines, implying varying shades of ambiguity based on the interpretation of the 'grey' element (Roth 2010). Indeed, as Schöpfel (2011) notes, the pivotal work on grey literature — Charles Auger's (1975) *Use of Reports Literature* — somewhat ironically does not actually include the term once. It is only with the later publication *Information sources in grey literature* (Auger 1989) that the phrase is coined: describing a range of elusive and often temporary materials such as conference proceedings and speeches that were not to be found through traditional publishers, yet still part of the academic communication medium. Indeed grey seems both to pertain to its liminal publication status as well as its intellectual appeal to the brains grey matter (Mason n.d.), and the academic advantages of grey over 'white' publication such as "greater speed, greater flexibility and the opportunity to go into considerable detail" (Auger 1989: 3).

The term soon crossed over into the archaeological sphere, as Seymour (2010b) notes, it is referenced within Cultural Resource Management documents in the United States in the early 1990s. Across the Atlantic, the exact point at which 'grey literature' became used in the UK to replace existing terms such as 'archive' reports is somewhat opaque. The distinctive nomenclature is not used within any of the contemporary reports on publication and dissemination strategies released in the 1990s (Carver et al 1992; English Heritage 1991), or the initial survey of Research Frameworks by English Heritage (Olivier 1996). It may however have its roots in York, being used, within quotation marks, in the editorial of the very first issue of *Internet Archaeology* to denote the hard-copy reports produced as part of the split between archive and publication (Vince 1996). A contemporary review of reporting procedures in England, although not using the term 'grey', describes thus:

"... the circulation of such reports is very poor, rendering archaeological information inaccessible and limiting opportunities for peer review ...assessment reports lacked the sort of basic information that would make cataloguing and indexing easy. Even where reports were deposited in public records, retrieval was often difficult. In addition, the format of reports varied considerably and frequently lacked useful information such as source lists and copies of the brief and/or specification which informed the assessment work." (English Heritage 1995: 16).

It seems that the term was not yet ubiquitous, 'grey literature' does not appear within the important overview of investigative trends based on analysis of unpublished fieldwork reports of the 1990s (Darvill and Russell 2002). However, it had gained enough recognition by the turn of the millennium to be defined in the Council for British Archaeology's report (also published in York) on archaeological publication: "those reports that are not issued for public sale or widespread distribution" (Jones et al. 2001). It is interesting to note that the same publication identified a marked ignorance within the community about what actually constitutes grey literature, as well as unawareness as to the extent of the genre (*ibid*). Despite, or perhaps because of this ambiguity, the term gained prominence and through works such as Bradley's (2006) seminal paper on the split between research and developer funded fieldwork becomes synonymous with the range of reports created through commercial investigations. Through this association, as grey has become embedded in the vernacular the term has an increasingly been used as an indicator of not only the unpublished nature, but also of perceptions of quality and availability (Aitchison 2010: 237; Chadwick 1998).

2.2. Modern grey literature

It is pertinent to dwell on these archaeological experiences and interpretations, as outside the discipline approaches to grey literature have continued to evolve (Schöpfel and Farace 2010). As previously noted, to the community that arose following Auger's publications, grey did not imply any assessment on quality of content but a characterization of the non-commercial distribution mode (Mackenzie Owen 1997). Indeed, grey continues to be championed as a complement to traditional publication outlets across many disciplines (Seymour 2010b). However in recent years increased access to grey literature via web-based dissemination, coupled with Open Access to

previously subscription-based journals has led to levels of 'grey' becoming blurred to the point of an existential uncertainty in the validity of the term itself (Schöpfel 2011). In addition the concept of what constitutes the genre is constantly being re-evaluated, with one survey identifying over 100 different types including websites and datasets within the classification (Pejsova and Vaska 2011: 10). Others have pushed this further to include the growing levels of online content such as blogs as increasingly relevant to both discipline specific meta-analyses and scientific synthesis (Banks 2012; Rothstein and Hopewell 2009). Prompted by the need to (re)establish what is grey in the modern world, the term is currently nearing its third official incarnation, with the most recent (proposed) Prague definition suggesting that it:

".. stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body." (Schöpfel 2011).

The inferences for archaeology are clear; what we consider our 'grey literature' is but a drop in the wider semantic ocean. Indeed it is illuminating to compare archaeological perceptions to the classifications of those outside the discipline. A search on *OpenGrey* — the database of 700,000+ bibliographical references of paper grey literature produced in Europe — reveals some diverse archaeological sources catalogued by the British Library, for example:

- A scientific dating report 2002 produced by English Heritage
<http://hdl.handle.net/10068/457603>
- A PhD thesis <http://hdl.handle.net/10068/392747>
- English Heritage Archaeological Report (ISBN 1-85074-681-8) of the excavations at Etton, Cambridgeshire: <http://hdl.handle.net/10068/397382>
- Proceedings of the 1990 Computer Applications and Quantitative Conference, published in the BAR International Series (ISBN 0860547132). <http://hdl.handle.net/10068/514794>

The last two 'published' sources in this list would strike us as contentious; however, as neither is controlled by major commercial publishers, they are technically 'grey literature'. Indeed, a report from the British Library shows the extent to which grey is very much in the eye of the beholder, with familiarity or ignorance of the resource dictating opinion as to its status (Tillett and Newbold 2006: 71). Thus it seems that for quite some time, the archaeological community has been producing a wide range of grey materials aside from developer-funded unpublished fieldwork reports (c.f. GreyNet n.d.). Examples include documents produced or funded by national organisations (e.g. <http://dx.doi.org/10.5284/1024571> Landward Research Ltd 2014), local societies/community groups (e.g. <http://dx.doi.org/10.5284/1018067> Amadio et al. 2007), or reports on grey literature itself (e.g. <http://dx.doi.org/10.5284/1012487> Museum of London Archaeology 2012). In addition to these reports, there is also the wide range and large number of online sources that increasingly play a prominent role in the dissemination of information and opinion (Jeffrey 2012). Clearly, the reality of archaeological 'grey material' is far larger and more diverse than traditionally envisaged.

Where then does this leave those documents generated through development-led investigations? In my opinion, whilst they are clearly a *type* of grey literature as covered by the Prague proposal, the application of this term in respect to reports from commercial investigations has perhaps outlived any original usefulness. Not only is it semantically limited, but is also a disciplinary relic that, while once understandable in wake of the publication crises in rescue and PPG16 funded projects, has become a cultural misnomer. When the term entered the lexicon in the mid-1990s it was in relation to the accessibility of hard-copy reports within research archives, produced to much debated criteria and against a background of publication crises (Jones et al 2001: 2.3.2-2.3.5; Richards and Hardman 2008: 27-28). However, nearly 20 years since that point - and 25 since the advent of PPG16 - there is perhaps less unease about the 'unknown unknowns' lurking in the disciplinary closet. The Archaeological Investigations Project (AIP) and OASIS have done much to catalogue and disseminate reports to the obvious benefit of the research community (Darvill and Russell 2002; Hardman 2011). In addition, by referring to information within the reports as 'grey', we are also ignoring the role it already plays within informing the Historic Environment Records of the country (Newman 2010: 2). Indeed, to label these documents as 'grey literature', as understood in its traditional archaeological sense, is to

ignore the fact that through HER database records, as well as access to physical holdings themselves, the information from these reports have always been accessible.

It may also be argued that the practice and reporting of fieldwork has changed to such an extent that historic perceptions and classification of grey literature are defunct. This is in the main dictated by the nature of archaeological fieldwork, which since the rise of smaller assessments in the 1980s has become dominated by evaluations and watching briefs, rather than full-scale excavation which may be expected to produce a 'research' archive or an interim/post-excavation report of significant value (English Heritage 1995; Darvill and Russell 2002). Even a very broad overview of more recent statistics shows the extent to which this pattern has become embedded within the country since the advent of PPG16 in 1990 (Figure 1). Indeed, it is also worth highlighting that the levels of partial/open excavation that may investigate larger areas or produce more results are not as large as may be imagined. For example, it is clear from the available statistics that levels of investigation are invariably cyclical reflecting not only planning policy but also the impact of contemporary economics on the levels of development occurring in the country. However, when looking at the relative levels of evaluations/watching briefs and excavations post-2000 it is clear that while the former plateaus, the latter group seemingly drops to their lowest level since the 1960s (Figure 1 and 2). On one hand, this may be a consequence of limitations in the coverage of the Excavation Index dataset, or perhaps the classification of 'evaluation' and 'excavation' (see Section 1.2). It is arguable however that the downward trend is not just a consequence of the data, but reflective of a very real trend. If so, these figures may be a consequence of a larger-more coordinated scale than of previous decades, but also of the decision *not* to undertake excavations as part of the mitigation for development. It certainly suggests however that the vast majority of the reports produced by commercial archaeology are simply for fulfilling the role of the fieldwork event to inform the archaeological mitigation within the final planning decision, not of 'interim' reports of larger works. Although all written materials are obviously of use to the researcher, it is debatable as to whether the majority of documents from these smaller investigations are really the type of hidden yet information-rich resources of popular imagination (see also Holbrook and Morton 2008: 56; Fitzpatrick 2012: 150). As Bradley (2006) noted in his research, the collation of nearly 20 years' worth of new discoveries from unpublished reports was undertaken by two individuals.

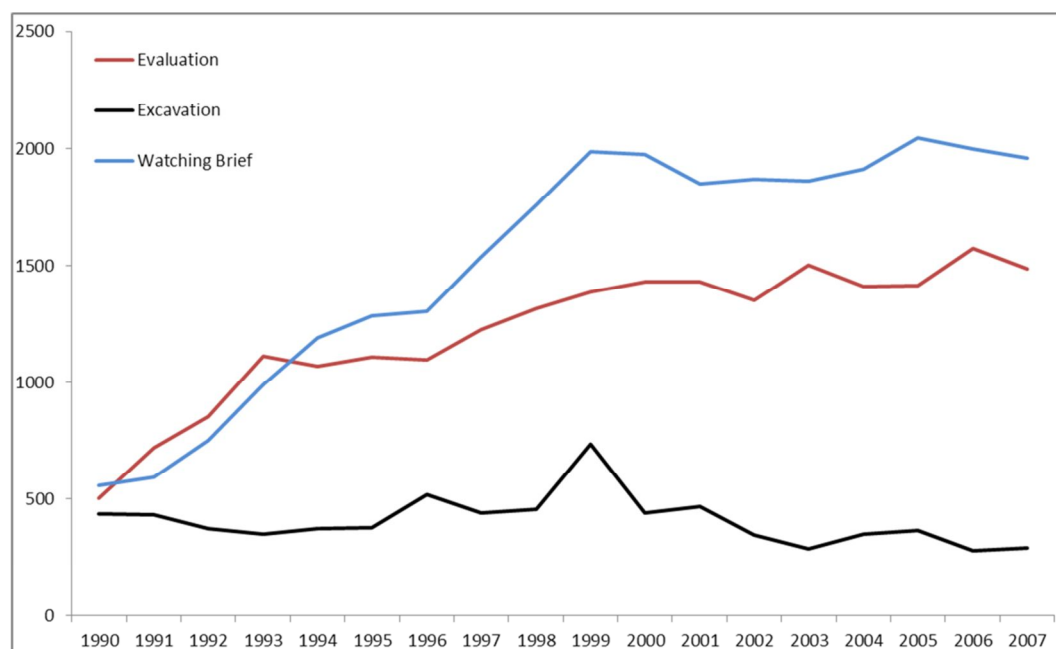


Figure 1: Development-led excavations, evaluations, watching briefs in England 1990-2007 as recorded in the Historic England Excavation Index. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014].

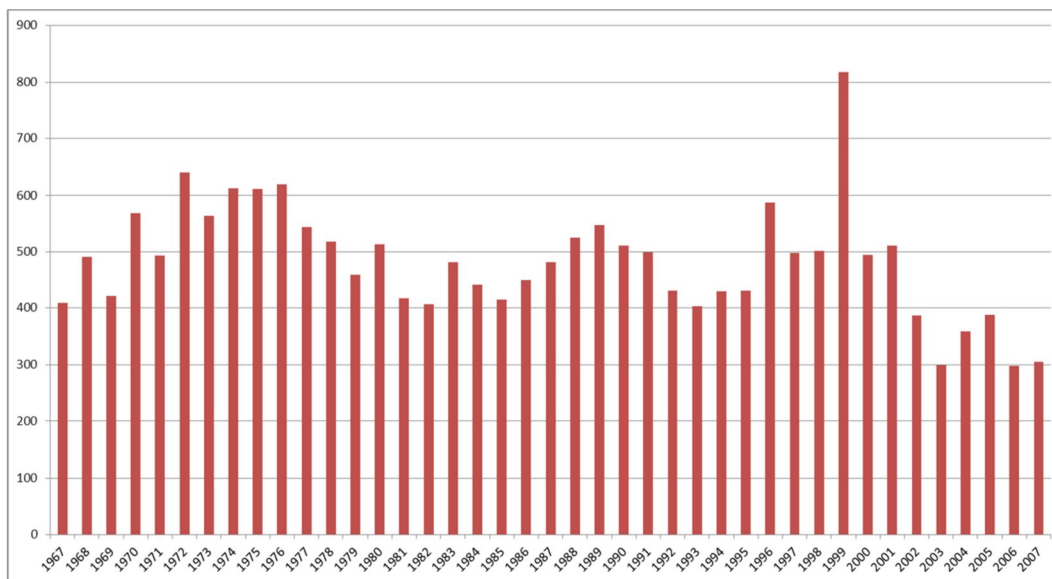


Figure 2: All excavations in England 1967-2007 as recorded in the Historic England Excavation Index. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014].

In addition, it is also clear that traditional disciplinary notions of both accessibility to, and the role of the fieldwork report, have changed irrevocably with the advent of online dissemination. For example, even a cursory search of the ADS Library of Unpublished Fieldwork Reports shows the extent to which online documents have become a valid route for the dissemination of results from both smaller evaluative and larger post-determination excavations. Increasingly it seems that for smaller excavations with results that do not require an Updated Project Design or a larger publication format, online 'assessment', 'archive' or 'data structure' reports appear to have become a de-facto publication outlet across the country, for example:

- Cornwall (Goacher and Mossop 2011)
- Cumbria (Hunter Blair 2009)
- Worcestershire (Rogers 2010)
- Leicestershire (Morris 2012)
- Suffolk (Sommers 2013)
- Greater London (Barrowman 2012)
- Northumberland (Amat and Cockburn 2009)

All of these reports are produced electronically and specifically for wider online dissemination (via the OASIS system) as part of the reporting/archive strategy: effectively bypassing the old accessibility issues and altogether a different phenomenon to their typescript predecessors. Indeed, it is arguable that the digital reports produced today have little in common with their counterparts from the early 1990s. As documents have become produced and disseminated digitally, presentation and content has evolved accordingly. For example, consider three excavation reports from the county of Lincolnshire and three reports by one organisation produced over a 20 year period:

- A report written 1990 by the Trust for Lincolnshire Archaeology on excavations at Corporation Yard, Boston Lincolnshire (Trimble and Brown 1990)
- A report written in 1990 by Wessex Archaeology on excavations at Montefiore Halls of Residence, Southampton. (Crockett 1994)
- A report written in 2000 by Pre-Construct Archaeology on excavations at Nelson Road, Fiskerton, Lincolnshire (Anon 2000)
- A report written in 2000 by Wessex Archaeology on work at Addington Street, Waterloo, London (Wessex Archaeology 2004)
- A report written in 2010 by Cambridge Archaeological Unit on excavations at Langtoft, Lincolnshire (Hutton and Dickens 2010)
- A report written in 2010 by Wessex Archaeology on investigations at Seven Stars Public House, Marsh Baldon, Oxfordshire. (Good 2014)

Comparing these reports it is clear that, leaving aside the thorny and subjective concepts of quality of interpretation, that as time progresses reports become relatively standardised, with very clear and detailed information on the site location, project codes (including museum accession numbers), methodologies, site plans, specialist finds reports abstract of results and so forth.

The improvement in the quality of fieldwork reporting is corroborated by the results of an English Heritage funded survey into the attitudes of the (UK) archaeological profession to 'grey literature' (GLADE), which in 2010 found that the majority of correspondents thought that the quality of information within fieldwork reports was satisfactory (Hardman and Evans 2010: section 4). The most common criticisms were of signposting the location of the archive and the interpretation of results especially in respect to the wider archaeological context; although it should be noted that as many correspondents that felt that the interpretation was 'poor' thought that it was 'good' (*ibid*; Figure 3). The same survey also identified an interesting dichotomy to attitudes to the grey literature' corpus, one hand those who thought it problematic:

- "There is a problem of access and awareness of its existence".
 - "Knowing it exists and getting hold of it can be difficult" (*ibid*: Appendix 1).
- On the other, those who did not view these reports as an issue of concern:
- "... it is the fulfilment of an obligation. It seems to be mainly academics who are having a problem in coping with too much primary data."
 - "... unavoidable as long as most work is aimed at the planning process: very little of it deserves/needs to be other than grey lit. It is a problem when a site that deserves to be published languishes as grey lit. but there probably aren't as many of those as people might imagine/fear. (sic)"
 - "Providing they are made accessible then there is no problem as it is unfeasible to publish everything" (*ibid*: Appendix 1).

These findings demonstrate a fascinating polarisation of views on access, undoubtedly dictated by regional experiences and sectorial perspectives (see Aitchison 2010). What they do suggest is that — at the time the GLADE report was written — a 'grey' report had become a valid method of publication in the literal sense of the term and through online access achieving a sense of immediacy in the dissemination of results. Indeed, with increased access via the web it is tempting to look back to Auger's original focus on the relative benefits of 'non-white' publication, especially in contrast to the results of projects that may take longer to publish via traditional hard copy methods (Holbrook and Morton 2008: 41). This is not of course to advocate that these publication methods be replaced, but rather that 'grey' reports are a fundamental facet of the dissemination strategies of modern archaeological practice. Furthermore, as the GLADE results show these are fit-for-purpose documents that, through the checking and validation by Local Authority archaeologists and consultants that is inherent in the planning process, are effectively 'peer reviewed'. Yet this cautious optimism may also be balanced by the increased pressure on local government archaeologists in recent years (ALGAO 2013). This not only affects involvement in the planning process itself, but also the capacity for Local Authority archaeologists to check and comment upon reports (Rescue 2013). The ramifications for this on the outputs produced through commercial archaeology are serious; although there may always have been a difference in reporting strategies according to the requirements of Local Authorities, the quality may be impacted by the lack of this 'peer review'. Indeed, from a point where increased use of web technologies has potentially created a new and interesting life for grey literature, the old fashioned factor of finance, and the choices of Local Authorities in their choice of funding and structure, is potentially restrictive.

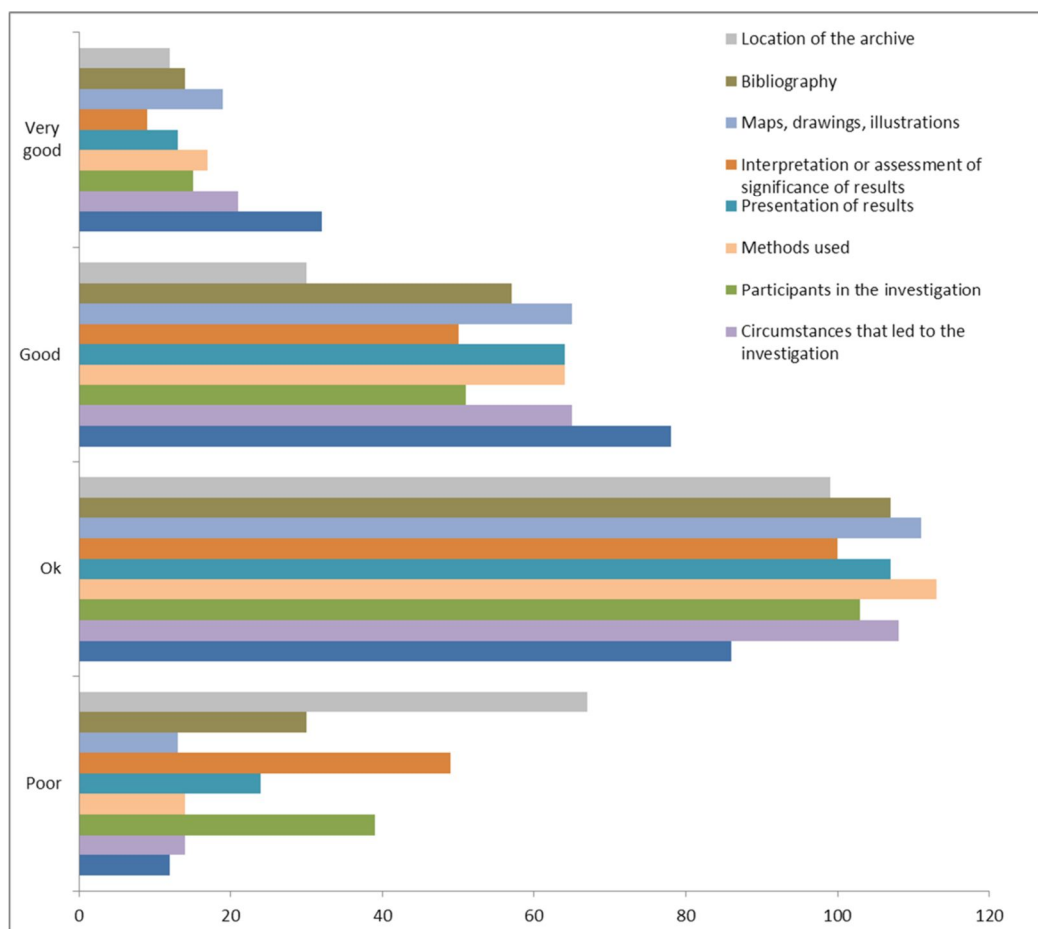


Figure 3: Perception of quality of information within fieldwork reports as reported in the 2010 GLADE survey (after Hardman and Evans 2010, fig. 4.6.3).

3. Problems and paradoxes

3.1 The nature of modern investigation

In the 25 years since the advent of PPG16 in England, the benefits and tribulations of grey literature as a data source for academics are well documented; namely the rapid rise of development-led works and the accompanying issues of accessing and collating the results (see Fulford and Holbrook 2011). As attitudes have begun to change, unpublished fieldwork outputs have been championed as valuable resources providing insight into geographic areas outside the zones of traditional research interest (Vander Linden and Webley 2012). Thus, in England, the opportunity for national syntheses on a scale previously unimaginable, or achievable, are now being taken by a number of projects. However, there are potential weaknesses inherent in cultural collecting patterns based on modern development that are perhaps less well studied (see also Last 2012: 136). This is somewhat surprising, as national projects and sources such as AIP, OASIS and the Excavation Index have done much to compile and catalogue archaeological investigations and outputs (Evans 2013).

For example, a broad analysis of the available statistics shows that intrusive planning-led events undertaken since the advent of PPG16 have a marked unevenness in distribution (Figure 4). If these numbers are restricted to just excavations — as opposed to evaluations or watching briefs — the distribution is even more skewed towards a core zone in the south of the country (Figure 5). This pattern is due to a combination of cultural and environmental factors underpinning the nature of modern development, and evident in the regional differences in modern planning applications and their potential impact on the historic environment (Table 1). This trend has been highlighted to some extent since the first analyses of development-led statistics (English Heritage 1995; Darvill and Russell 2002), but whilst noted in certain studies (Holbrook and Morton 2008; Morrison et al 2014) there has generally been a lack of any analysis of the significance of these

cultural collecting patterns. This is arguably an important oversight, as the disparity in twenty years' worth of investigations has arguably created zones of core and periphery within the modern archaeological landscape. For example if one wanted to examine 'Iron Age' sites encountered by commercial archaeology the resulting dataset shows a distinct geographical clustering (Figure 6). Thus, on a national level the levels of data, including unpublished reports, being produced are a palimpsest of the modern planning landscape, rather than perhaps an accurate depiction of the archaeological resource itself. Although patterns of bias are nothing new in archaeology, in the current climate of large-scale synthesis it is imperative that they are highlighted and factored into models of past settlement. Indeed, the most interesting questions may not just be restricted to what the data tells us that we know, but what it also tells us about limitations and our own roles in creating datasets of cultural disparity (see also Cooper and Green 2015)

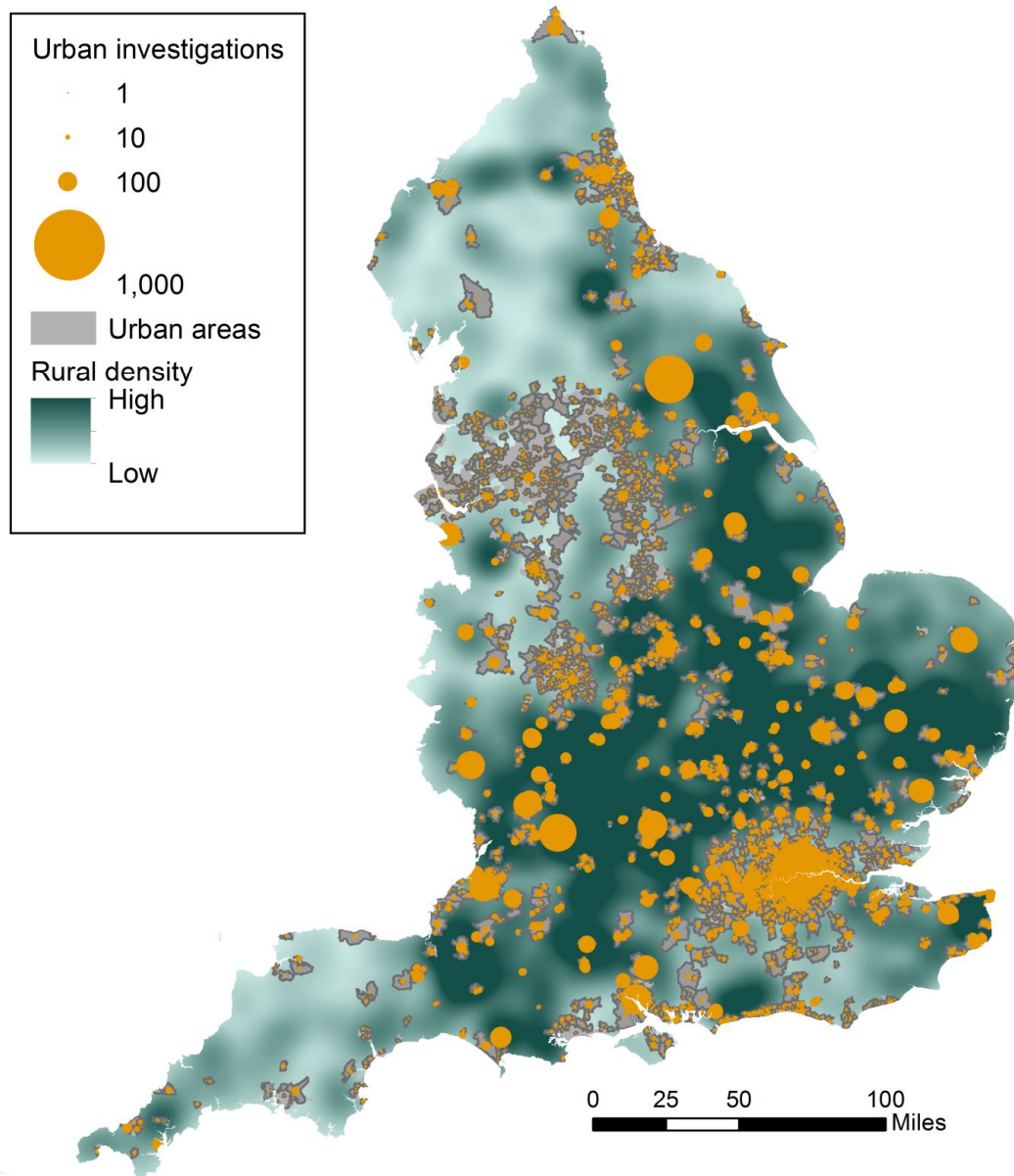


Figure 4: Development-led rural and urban excavations, evaluations and watching briefs in England 1990-2010. Rural investigations displayed as kernel density (20km); urban investigations displayed as a proportional symbol for each urban Middle Super Output Area (MSOA). Rural/urban definition provided by the 'Rural-urban classification (2011) of middle layer super output areas (2011)' (.zip) available for use under Open Government Licence terms and conditions. [Last accessed: 23 Oct 2014].

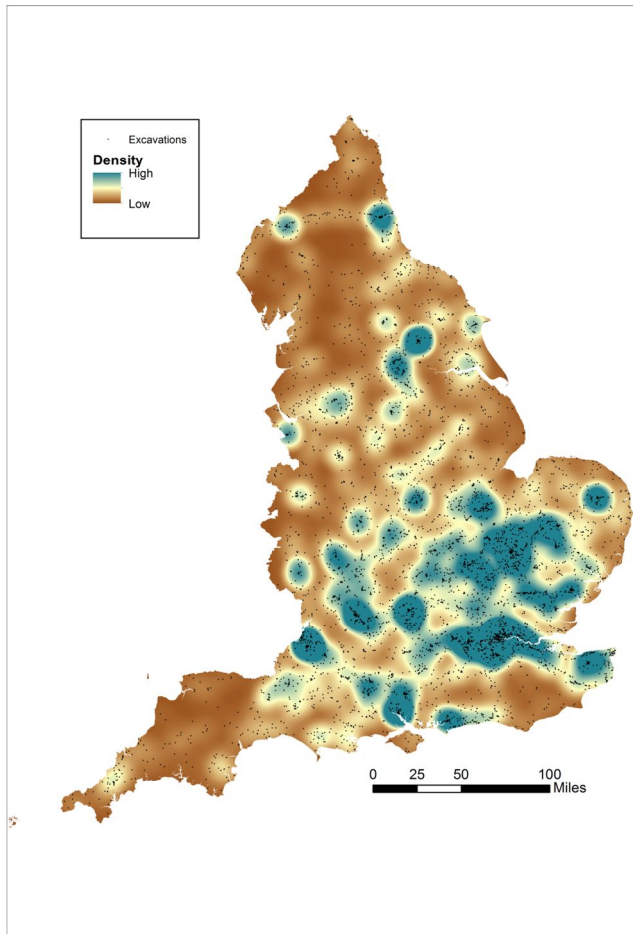


Figure 5: Kernel density (20km) of development-led excavations 1990-2010. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014].

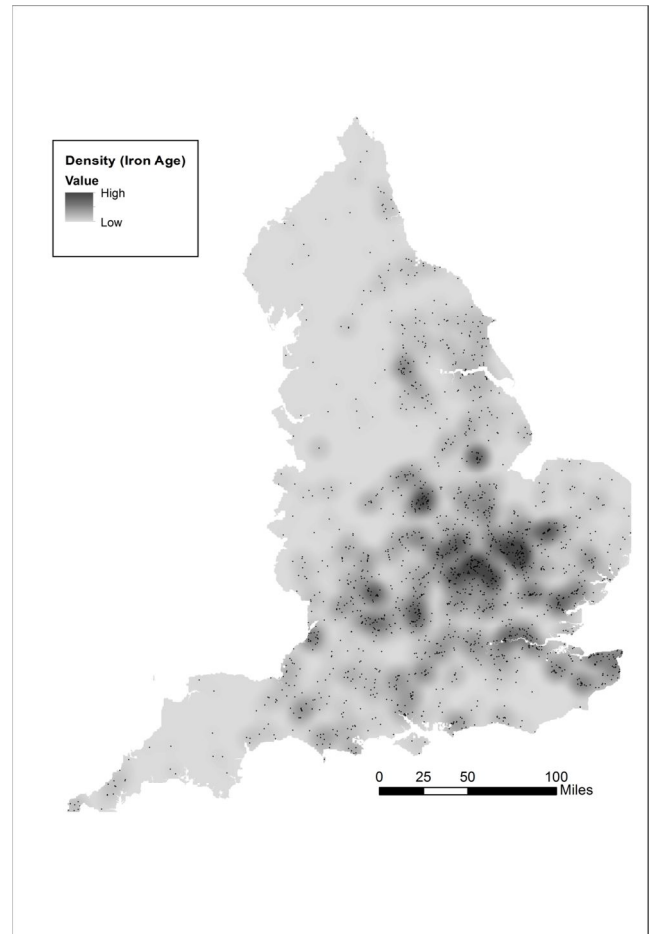


Figure 6: Kernel density (20km) of development-led excavations 1990-2010 encountering deposits of 'Iron Age' date. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014].

Table 1: Planning applications in England 2003-2011. Data from ALGAO: England

Planning applications	2003	2004	2005	2006	2007-8	2010-11
East of England	70978	76958	73973	44471	54898	66138
East Midlands	45214	35506	39340	45266	30489	42928
London	97870	104542	96977	89262	99394	114969
North East	26523	28635	28118	26978	12000	20374
North West	68610	83261	60398	49383	37211	48079
South East	96246	109317	59340	60716	84839	82067
South West	65832	41797	84309	83722	56394	88464
West Midlands	32392	39773	41352	21934	25915	27962
Yorkshire & Humber	59051	63098	62421	59249	53238	46973
England Total	562716	582887	546228	480981	454378	537954

It may be argued that despite these national statistical discrepancies the advent of developer engendered works still represent a distinct cultural break at a more local level. For example, looking at just two English counties – Staffordshire and North Yorkshire, as representing relatively quiet and busy archaeological areas respectively – it is evident that the advent of PPG16 represents a clear statistical leap from previous decades (Figures 7 and 8). However, it may also be observed that although the trend is for more investigations, these are increasingly geographically clustered around areas with higher levels of development, communication routes or areas of aggregate extraction; a consequence of the inexorability nature of threat and curatorial knowledge (Roskams and Whyman 2007; Powlesland 2011). Such criticisms of development or rescue led archaeology are of course not new, since the first rescue excavations many authors have highlighted the uneven distribution of funds, and projects (Fowler 1977; Biddle 1994). However, the primary challenge over the last few decades has been one of collation and broad categorisation, and although some detailed work has been undertaken primarily through the *Assessment of Assements* and AIP, there has been a distinct lack of follow up on this work by the wider discipline as a whole (cf. Darvill et al 1995; Darvill and Russell 2002). While we rightly celebrate an increased capacity to collate and synthesise, an increased predictability in where investigations have been, and are going to be, undertaken presents an interesting theoretical dilemma regarding the inherent entropy of the unpublished reports generated by so many projects; that is the *value* of data in contrast to the whole (Justeson 1973: 132). As more of the same types of sites in the same areas are being excavated, does the information collected within fieldwork reports present new knowledge, or just more data?

As an aside, it is interesting to note the extent to which this distribution compares to those intrusive events undertaken outside of the planning system: ‘research’ excavations by Universities and Local Societies and other community groups or researchers. Although to a large extent these mirror the broad southern-centric patterns of investigation, there are notable divergences, particularly with the concentration of academic excavations in the central-southwest region (Dorset and Wiltshire) and the significant clusters of works undertaken by local groups in the areas to the south of London (Figure 9). For this admittedly very broad picture it seems that what may very loosely be termed cultures of ‘research’ and ‘mitigation’ (see Fitzpatrick 2012) have become physically disparate. In addition, it is notable that whilst the numbers of works being undertaken by these groups is significantly smaller, they are not insubstantial. In particular community groups are responsible for a perhaps surprising volume of work, according to the Excavation Index over 1000 intrusive investigations 1990-2010. Of course the relative size of these investigations should be taken into account, multi-period large scale works versus short bursts of work by relatively small numbers of people. Issues of scale and significance apart, there are clearly significant important levels of work, produced outside of contract archaeology (see also Holbrook and Morton 2008). Conversely, it may be suggested that whereas rescue/development-led works were once highlighted as counter-balancing the monument based biases of previous research, it seems that the activities of modern community based research groups may present a geographical alternative to commercial archaeology. However, as previous studies have shown, it is these groups rather than contracting units that are under-represented in the collections of HERs and national databases of AIP and OASIS (Brookes 2003; Evans 2013). Thus, it may be that the real ‘grey literature’, elusive and ephemeral, may well be those written and produced outside of the two cultures of academia and development control.

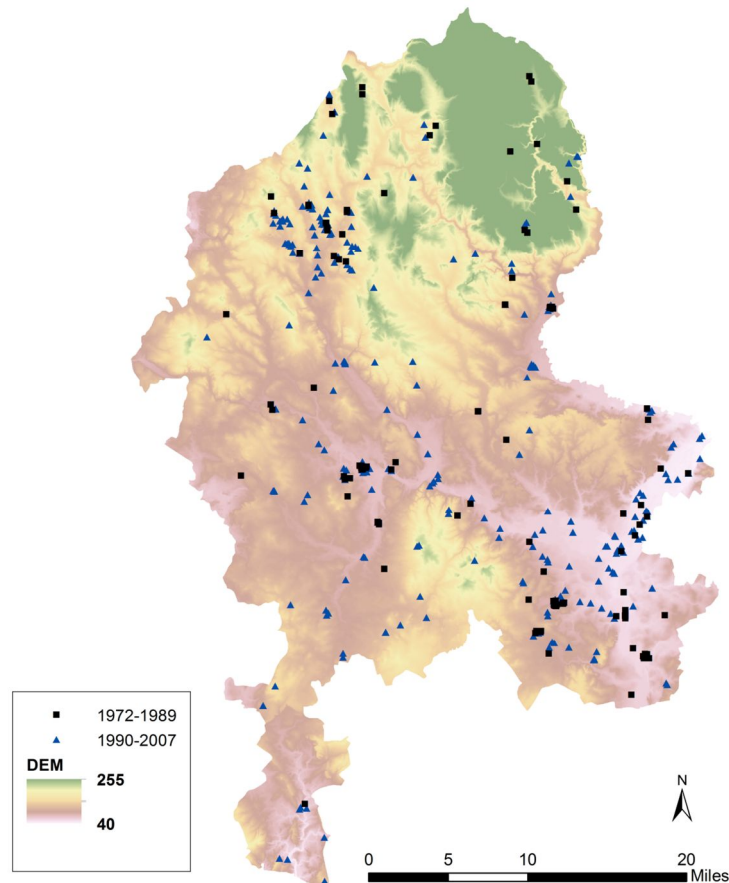


Figure 7: Development-led intrusive investigations 1990-2007 compared to all intrusive investigations 1972-1989 in Staffordshire. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014]. Height data derived from OS Terrain 50 DTM © Crown Copyright and Database Right Ordnance Survey (Digimap Licence).

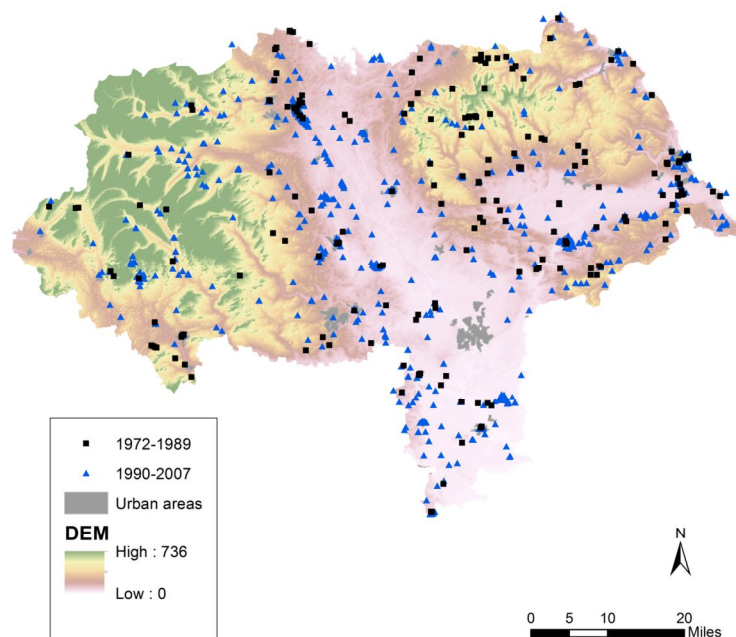


Figure 8: Development-led intrusive investigations 1990-2007 compared to all intrusive investigations 1972-1989 in North Yorkshire. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed: 1 Sept 2014]. Height data derived from OS Terrain Terrain 50 DTM © Crown Copyright and Database Right Ordnance Survey (Digimap Licence).

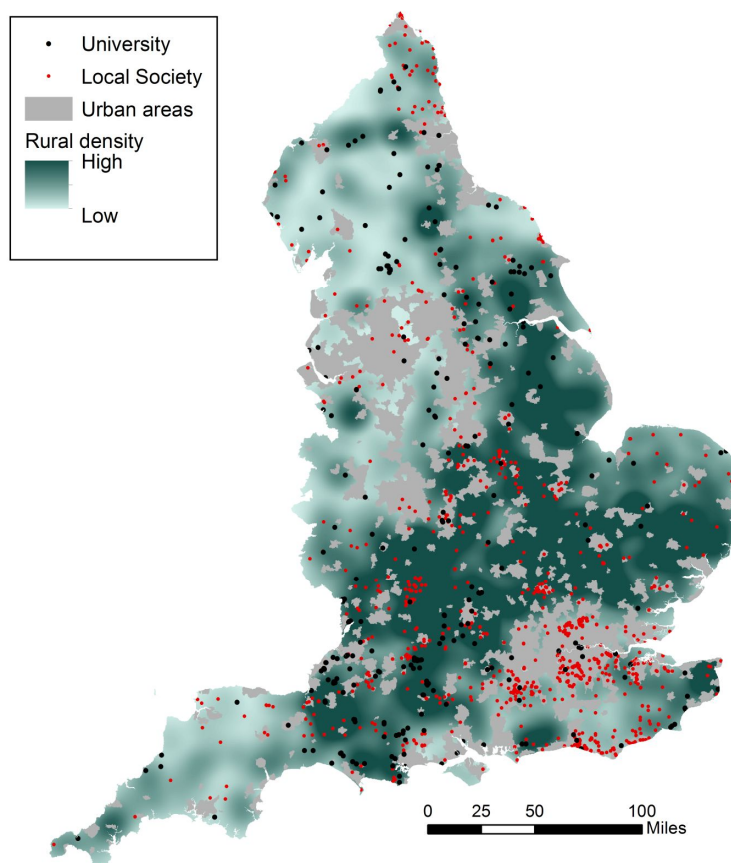


Figure 9: University (n=322) and Community Group (n=1126) intrusive investigations in England 1990-2010, overlying a kernel density (20km) of development-led excavations 1990-2010. Data from the Excavation Index hosted by the Archaeology Data Service [Last accessed : 1 Sept 2014].

3.2 The nature of dissemination

Despite the clear biases displayed, this is not to say however that synthesis and interpretation of the data produced through commercial (and indeed other sources) of investigations are invalid; clearly, the results from past and present studies are essential (see Bradley 2007). However, the spectre of sampling bias should not be ignored but rather met head-on, and undoubtedly the role of the new generation of academic syntheses should be to engage with the gaps to identify what we don't know, as well as what we do. Indeed, these gaps are arguably the final frontier for new research undertaken by the academic and community sectors. However, in order for this to happen it is imperative that the results of these syntheses and analyses, including lacunae, are disseminated as widely as possible: a rallying call to challenge the potential known unknowns still within the landscape (c.f. Garwood 2007). Indeed, the potential paradox in the increased re-use of grey literature by academia is that freely available information (i.e. fieldwork reports) are collated and absorbed solely into traditional hard copy formats that are themselves not widely accessible. For example, the landmark works on later prehistory and historical periods that leant heavily on commercial archaeology, and cited by Aitchison (2010) as examples of successful re-use both had published in this form (Bradley 2007; Newman et al 2001). It is debatable as to the extent to which these (commercial) outputs are readily accessible to those without access to a well-stocked University library: specifically curators and units responsible for the original data on which these syntheses are based, or locally based groups wishing to engage in research in their geographic region.

It should be noted however that this is not a direct criticism of these particular projects or the publication strategy itself; the incentive for Higher Education-based researchers is to publish, especially in monograph form in order to demonstrate research excellence and impact (OAPEN-UK 2014). Open Access journals are of course one method, but are currently not widespread within the major period-based journals within the UK. Indeed, the journal format may not be

conducive to the publication of national period-based overview that encompasses large volumes of text and graphics such as Bradley's (2007). Although there are steps in Higher Education publishing towards Open Access monographs this is not yet widespread, and not yet a requirement within UK Higher Education funding bodies (Crossick 2015). Thus, traditional publication creates a potentially paradoxical situation: when academic synthesis is undertaken, it is in danger of becoming relatively 'grey' itself: not widely accessible outside of particular institutional collections and the culture of academia. In addition, it has also been noted that publications, unpublished reports and data from academics are often missing (i.e. not submitted) to HERs, thus perpetuating the cultural and methodological schism (Brookes 2003; Evans 2013). Conversely, if the academia do not collate, analyse, synthesise and publish then these tasks are — large infrastructure schemes such as Terminal 5 or CTRL aside — not carried out within the typical commercial framework (Morrison et al 2014).

The answer to this potential impasse appears in the online strategies of two recent projects: foremost, the Roman rural landscape project whose publication strategy incorporates access to the research dataset, itself a research-driven appraisal of the value of unpublished fieldwork reports (Smith pers comm). The second solution appears to lie within the new generation of Research Frameworks as outlined by English Heritage (Miles 2013) and exemplified in the recent East Midlands document being made available online as an updatable wiki (East Midlands Heritage 2015). Through having flexible and free to access documents there is the real potential for regions (and periods) to have up to date and comprehensive research priorities which could be used to inform and influence the nature planning/developer-led fieldwork. Indeed, this could potentially put the emphasis back on *research* in the curatorial and commercial sectors, enabling developer funding to be used to help fill the research gaps instead solely having to rely on academics taking an interest and securing the budgets to do so. An extension to this strategy could be to link these flexible regional and thematic documents to online fieldwork reports as they are created; thus enabling contractors to identify key 'unpublished' works that relate to a particular research theme.

Thus an intrinsic part of our 'grey literature' future is the online dissemination of fieldwork reports: bypassing traditional problems with accessibility and potentially offering new avenues for linking (predominantly) commercially funded outputs with research. However, although the strengths of this are clear (Moore and Evans 2013), perhaps less so are the potential weaknesses. The first issue is the disparities in the current geographical distribution of online reports; the unevenness of the location of commercially-led investigations is, to a large, extent mirrored by the online corpus (Figure 10). In point of fact, it is even more skewed towards the densely investigated central/eastern region. Compare this with the scant information from the Northern counties, as well as gaps in densely investigated parts of the South, and there seems to be a serious imbalance in online information. This is of course a reflection of the relative levels of fieldwork but also the cultural working practices and the use of OASIS by relevant curators and units, since the piloting of the system in the East of England (Smith et al 2012). In addition, it may also be suggested that this disparity and use is influenced to some extent by the logistical and financial capabilities of HERs to process unpublished reports and OASIS backlogs at a time of serious cuts to staff and resources (Rescue 2013). In addition to this, what is striking is the under-representation of research projects undertaken by community groups, academic institutions and individual researchers (Figure 10). As noted above, the sector outside of commercially funded works is not inconsiderable, but appears to not be submitting unpublished reports through OASIS. This is perhaps unsurprising for projects that do not generate any such reports, or those with a traditional publication strategy, but still suggests a cultural imbalance in the online availability of information.

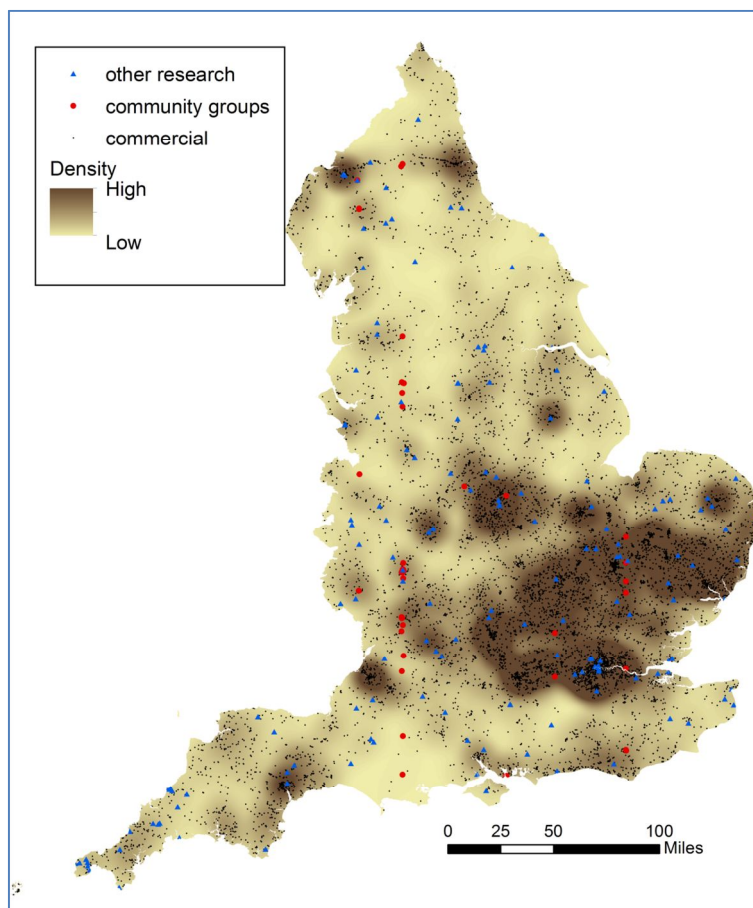


Figure 10: Reports for excavations, evaluations and watching briefs recorded in OASIS and disseminated via the Archaeology Data Service: for planning-led events (n=13978), investigations by community groups (n=169) and other 'research' projects (n=61). All records also displayed as kernel density (20km). Data from <http://archaeologydataservice.ac.uk/archives/view/greylit/> [Last accessed: 22 Oct 2014].

It may also be argued that as the number of online reports grows, the nature of accessibility has changed. As Huggett (2014) has recently outlined there are a raft of practical and theoretical issues concerning the limitations of large-scale data dissemination. To this, we may add the discoverability and usability of these reports files themselves. Outside of archaeology, a recent survey of reports held by the World Bank shows that simple accumulation of PDFs with basic bibliographic metadata is not conducive to re-use, but does in fact create vast online 'graveyards' of un-used information that is accessible in name only (Doemeland and Trevino 2014). It should be stated that this is clearly not the case for archaeology, as access statistics and objective assessment clearly show the use and impact of online reports (Beagrie and Houghton 2013). However, looking forward there is, given that the online corpus is growing at an increasing rate, a need to re-visit and explore new notions of access and discoverability, and not simply relying on the fallacy that more data equates to more knowledge. As the recent projects using unpublished reports for research into the Roman period has shown, a great deal of time can be spent establishing not only the extent of the resource, but also its usefulness (Holbrook and Morton 2008; Fulford and Holbrook 2011). For example, at the time of writing the ADS library is exceeding 30,000 records, and a search for 'Medieval' 'England' returns over 6000 reports, 'Iron Age' returns over 1100 and 'Post Medieval' almost 10,000, with no further way to refine these queries except on the subjective recording of monument and artefact terms. It may be argued that this is enough, but while commendable does not yet meet the vision of online publishing and communication at the beginning of the current century:

"Whereas previous generations would have taken years to gather all the references to research at a given location and would have congratulated themselves on the achievement of such a laborious task, in the future (and increasingly in the present) this will be the work of a

single day or less. Comprehensive referencing will not be a virtue to which scholars aspire, it will be a *sine qua non*. It will no longer be the first stage of a doctoral thesis; it is likely to be the first stop in an undergraduate paper. The days of the descriptive index submitted without interpretive scrutiny are surely numbered" (Clarke et al 2003).

It may be argued that for meaningful and innovative searches to be undertaken underlying systems have to be advanced, less we returning to the simple accumulation of data criticised in the past (Thomas 1991). One answer to more intelligent searching of reports could well be machine-based, as demonstrated by recent research into the generation of rich metadata via text recognition (Vlachidis et al 2013). Although increased levels of metadata can be advantageous for locating specific monuments or artefacts, it may be of limited use to know that a report contains reference to 'Animal Remains' or 'Sherd', it may be of greater significance to know whether these are discussed in detail (and indeed by whom), and their relevance to a particular period or area of research. Archaeological reports are not databases or inventories, which as noted elsewhere may well be overlooked facets of the digital archive (Evans and Moore 2014), and their value is undoubtedly as literature no matter how 'grey' this is perceived. In tandem with technical innovation there is also need to remain grounded in the archaeological, context of these reports; not only 'who-what-where' but also 'why' and 'how'. At the time of writing detailed metadata about the context of the event such as prompt, land use development type, investigation size and methods employed are all recorded in the OASIS system but not replicated within the index of reports. The power of this potential metadata cannot be understated, and with the future redevelopment of the OASIS system and the ADS library it is imperative that this data is both retained and utilised.

It is also important that the accessibility and discoverability of reports is not confined to systems such as the ADS Library, no matter how successful this endeavour. The aforementioned Research Frameworks are of course on method, but there are also a growing number of online databases – for example the Bibliographies of Medieval Pottery and Roman Pottery that offer the potential specialist appraisal of the contents of a report (MPRG 2011; SGRP 2014). Large-scale syntheses such as the Rural Settlement of Roman Britain (Allen et al. 2015) also offer in-depth analyses of particular facets of monuments and sites. Last, but by no means least, there are also the HERs: the first port of call for monuments in England. Although somewhat optimistic, it may be envisaged that these often disparate systems can be aligned, incorporating the detail recorded within HERs of events and sources with additional metadata recorded by finds specialists and academics. Such a vision of 'linked data' is of course nothing new (see Tudhope et al. 2011), but it is now arguable that with the growing number of online resources, and even with the limitations discussed above, online research must advance beyond a simple type and hope approach, thus truly making 'grey literature' an integrated part of the information landscape.

Conclusions

Perhaps inevitably, any discussion of the 'grey literature situation' has shades of optimism, concern, and ambiguity. Foremost, and perhaps controversially, I would suggest that the archaeological community begins to move away from the trend of describing the outputs of planning-led investigations as 'grey literature'. Although a *type* of grey literature as described in modern library definitions, the wider term is describing all materials not controlled by commercial publishers. These include outputs traditionally viewed by archaeologists as published and those not considered literature at all. Moreover, the term may be of more use to non-discipline specific curators and librarians rather than fieldworkers and academics. Although it may have some use in referring to the traditional published/non-published dichotomy in fieldwork archaeology, the latter is actually a multi-faceted corpus which represents different forms of document (and information) from different types of event: primarily a large volume of small reports simply reporting the presence or absence of archaeological deposits. These are not

'grey' as understood by the original use of the term by Alan Vince and others in reference to the research archive produced by larger excavations, but essential assessments fundamental to the planning process. Even for larger excavations, it seems that the advent of online dissemination has provided an alternative publication route that may not be journal or monograph based, but is still generally of a good quality and a product of a mature field discipline. In addition, semantic quibbles notwithstanding, the interpretation of the phrase still pertains to the more negative aspects of the idiom that are, in reality becoming redundant. Given these modern developments, the use of the term 'grey literature' seems more and more like the relic of an obsolete vernacular.

Furthermore, the challenge for those interested in 'grey literature', or whatever the preferred nomenclature, has begun to evolve from simple access to fieldwork reports and the amalgamation of the results into the research paradigm. Conversely, the opportunity for research could be the areas that have not been heavily investigated through the planning process; engaging, understanding and explaining the gaps in distribution maps and the potential sites not in any literature. As Last (2012) has previously described, a holistic research that can be carried out by all industries within the archaeological culture. However, the marked disparities in digital coverage, including those potential gaps created by community groups and academics that are under-represented in the online corpus, makes this vision far from straightforward. Indeed, if current trends continue the online information base may well become dangerously skewed, inadvertently presenting a distorted view of a dataset that is already subject to the biases of planning-led investigations. Thus rather than breaking new ground and moving away from traditional collecting patterns (Darvill and Russell 2002: 53), there is a potential danger of simply revisiting densely investigated areas and perpetuating biases that are now based on the visibility of online data, as opposed to the visibility of monuments.

This is assuming of course, that reports can be found efficiently, and to answer specific research questions. As online libraries such as the ADS continue to grow, we have to ensure that they are developed to suit a widening world of online resources. Simply cataloguing reports and isolating them from their wider (and more detailed) context while initially helpful, is arguably not embracing the potentials of online access. Whilst we should, unreservedly celebrate online dissemination and nationwide syntheses as advances from the bad old days of backlog and epistemological uncertainty, this is not a time to stand still. As the volume of online fieldwork reports increases, and as the web becomes an acceptable dissemination routes for many (but not all) excavations, researchers are going to be presented with more and more data. Future questions may well revolve not around the physical accessibility of reports, but rather which reports can help answer specific and – it is anticipated – increasingly specialised questions. Indeed, the old categories and perceptions of grey literature may well be becoming obsolete, we must be careful not to replace them with colourless online inventories that catalogue much but explain little.

At the time of writing the institutions and mechanisms for transferring and curating archaeological information are currently under review, or being re-evaluated in light of national strategies (Boldrini et al 2015; OASIS 2014). Initial signs are that existing mechanisms of accessing data remain relatively unchanged, that is that effort is made to emphasize the role of the HERs as the primary source of information on archaeological monument and event data (*ibid*). In particular it is notable that the OASIS system will potentially form a significant part of this mechanism, and thus it is imperative that the benefits of using the system, in tandem with redesigns to allow a greater level of participation, are championed. This is especially true of the research community, as noted throughout this piece and in other studies it is this body of fieldworkers that often go under-represented in recording systems (Evans 2013). In recent years there has been a noticeable increase in academic engagement with development-generated data, this should extend to engaging with the recording mechanisms themselves. Indeed, building upon the strengths of these systems it is possible to envisage a more integrated national system than currently available, with events and associated reports recorded in detail by HERs and, in

some cases, linked to the ADS archives via a Digital Object Identifier (Hardman 2011). Accordingly events and sources should be part of any future national data portal, querying not only monuments but their investigative history and the sources available. As the recent Rural Roman Settlement of Britain project has shown these are now the types of questions being asked by researchers (cf. Holbrook and Morton 2008; 10.5284/1030449). A capacity to allow a more sophisticated interrogation of the evidence base, alongside the evidence itself, should be a minimum requirement, and it is perhaps illustrative that the cursory analysis of investigative trends reported here is limited by the gaps in coverage of our current national systems. Such a potential system would enable future research to be based on initial analysis of the investigative and recorded sources – identifying areas or themes that were under-represented or corpuses of material that offered new perspective on particular issues. The current wave of syntheses have, understandably, been based on catching up with the data, the works of the next decade should use the data to push agendas forward.

Acknowledgements

I particularly wish to thank Gail Falkingham for insightful comments on an earlier version of this draft, as well as Jo Gilham, Catherine Hardman, Louisa Matthews, Ray Moore and Julian Richards for various discussions that have informed this piece. Additional thanks are due to the team of the Roman Rural Settlement of Britain Project, principally Martyn Allen, Tom Brindle, Neil Holbrook and Alex Smith, whose various views and research on 'grey literature' and excavation trends have helped inform sections of this work. Additional thanks are due to Mark Barratt and Martin Newman for clarification of the role of the Excavation Index. Any mistakes and oversights are of course my own. Finally, I wish to thank ALGAO England for the provision of their statistics for incorporation within the article. The publication of this article was supported by a Departmental Research Grant from the Department of Archaeology, University of York.

Bibliography

- Aitchison, K. 2010 'Grey Literature, Academic Engagement, and Preservation by Understanding', *Archaeologies: Journal of the World Archaeological Congress* 6(2), 289-300. <http://dx.doi.org/10.1007/s11759-010-9145-5>
- ALGAO 2013 A fifth report on Local Authority Staff Resources. Retrieved from <http://www.ihbc.org.uk/skills/resources/5th-rep-LAStaff.pdf> [accessed 01/06/2014].
- Amadio, L., Clarke, B. and Staff, C. 2007 *Report of Survey of Barn Ground Barn, Bottom Farm, Rodbourne, Malmesbury, Wiltshire, Devizes*: Wiltshire Archaeological and Natural History Society, Report No. 167.0901. <http://dx.doi.org/10.5284/1018067>
- Amat, D. and Cockburn, P. 2009 *An Archaeological Excavation on land adjacent to West Street, Norham, Northumberland*, Gateshead: Archaeological Research Services, Report 2009/66. <http://dx.doi.org/10.5284/1004391>
- Anon 2000 *Medieval Manor, Nelson Road, Fiskerton, Lincolnshire, Archaeological Excavation*, Pre-Construct Archaeology (Lincoln). <http://dx.doi.org/10.5284/1012785>
- Archaeology Data Service 2011 English Heritage NMR Excavation Index for England <http://archaeologydataservice.ac.uk/archives/view/304/> [accessed 2 September 2014]
- Ariadne 2014 Grey Literature Special Interest Group. Retrieved from <http://www.ariadne-infrastructure.eu/Community/Special-Interest-Groups/Grey-Literature> [accessed 30 September 2014].
- Auger, C.P. 1975 *Use of Reports Literature*. London: Butterworths.
- Auger, C.P. 1989 *Information Sources in Grey Literature (Second Edition)*. London: Bowker Saur.
- Banks, M. 2012 'Blog Posts and Tweets: The Next Frontier for Grey Literature', *The Grey Journal* 8(1), 54-59.
- Barrowman, S. 2012 *Land at the Former Forest Laboratories, Weir Road, Bexley, DA5 1LW - An Archaeological Excavation*, London: Pre-Construct Archaeology Ltd (London). <http://dx.doi.org/10.5284/1022099>
- Beagrie, C. and Houghton, J. 2013 The Value and Impact of the Archaeology Data Service. A study and methods for enhancing sustainability. Retrieved from http://archaeologydataservice.ac.uk/attach/impact/ADSReport_final.pdf [accessed 21 October 2014].
- Biddle, M. 1994 'What Future for British Archaeology?' *Paper presented at the Eighth IFA Archaeology in Britain Conference 1994*. Oxford: Oxbow Lecture 1.
- Boldrini, N., Webster, C., Minting, A., Flower, C., and Tait, G. 2015 Data Supply and Reconciliation between NRHE and HERs. End-of-Project Report. Retrieved from https://www.jiscmail.ac.uk/cgi-bin/filearea.cgi?LMGT1=HERFORUM&f=/Data_Supply_and_Reconciliation [accessed 06 May 2015].
- Bradley R. 2006 'Bridging the Two Cultures - Commercial Archaeology and the Study of Prehistoric Britain', *The Antiquaries Journal* 86, 1-13. <http://dx.doi.org/10.1017/S0003581500000032>
- Bradley, R. 2007 *The Prehistory of Britain and Ireland*. Cambridge: Cambridge University Press.
- Bradley, R., Haselgrove, C., Vander Linden, M. and Webley, L.J. 2015 *The later prehistory of Northwest Europe: the evidence of recent fieldwork*. Oxford: Oxford University Press.

- Brookes, S. 2003 Bibliographic Database of Archaeology in the Trent Valley. Project documentation and metadata. <http://dx.doi.org/10.5284/1000125>
- Carver, M.O.H., Chapman, H., Cunliffe, B., Hassall, T., Hebditch, M., Lawson, A., Longworth, I., Morris, R., Phillipson, D., Schofield, J. and Wainwright, G. 1992 *Archaeological publication, archives and collections: towards a national policy*. London: Society of Antiquaries of London.
- Chadwick, A. 1998 'Archaeology at the edge of chaos: further towards reflexive excavation methodologies', *Assemblage* 3. Retrieved from <http://www.shef.ac.uk/assem/3/3chad.htm> [accessed 17 September 2014].
- Clarke, J.P., Hardman, C.S. and Kilbridge W.G. 2003 'Comprehensiveness for All: The OASIS Project and Research Values in the Digital Age', CSA Newsletter 15(3). Retrieved from <http://www.csanet.org/newsletter/winter03/nlw0305.html> [accessed 28 January 2015]
- Cooper, A. and Green, C. 2015 'Embracing the Complexities of 'Big Data' in Archaeology: the Case of the English Landscape and Identities Project', *Journal of Archaeological Method and Theory* 22(1). <http://dx.doi.org/10.1007/s10816-015-9240-4>
- Crockett, A. 1994 *Excavations at Montefiore Halls of Residence, Swaythling, Southampton, 1992*, Salisbury: Wessex Archaeology. <http://dx.doi.org/10.5284/1027346>
- Crossick, G. 2015 Monographs and Open Access. A report to HEFCE. Higher Education Funding Council England. Retrieved from <http://www.hefce.ac.uk/pubs/rereports/year/2015/monographs/#d.en.99908> [accessed 28 January 2015]
- Darvill, T., Burrow, S. and Wildgust, D., 1995 *Planning for the Past. Volume 2: An assessment of archaeological assessments 1982-1991*. Bournemouth & London: Bournemouth University and English Heritage.
- Darvill and Russell 2002 *Archaeology after PPG16: Archaeological Investigations in England 1990-1999*. Bournemouth: University Press/English Heritage.
- Doemeland, D. and Trevino, J. 2014 Which World Bank Reports Are Widely Read? The World Bank Policy Research Working Paper 6851. Retrieved from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2014/05/01/000158349_20140501153249/Rendered/PDF/WPS6851.pdf [accessed 12 September 2014]
- East Midlands Heritage 2015 An Updated Research Agenda and Strategy for the Historic Environment of the East Midlands. Retrieved from <http://archaeologydataservice.ac.uk/researchframeworks/eastmidlands/wiki/> [accessed 12 December 2014]
- English Heritage 1995 *Planning for the Past Volume 1: a review of the archaeological assessment procedures in England 1982-91*. London: English Heritage.
- Evans, T.N.L. 2013 'Holes in the Archaeological Record? A Comparison of National Event Databases for the Historic Environment in England' *The Historic Environment* 4(1), 19-34.
- Evans, T.N.L. and Moore, R.H. 2014 'Beyond 'The Preserving machine': Promoting Use and Reuse of Grey Literature and Grey Data'. *The Archaeological Review from Cambridge* 29.2, 212-225.
- Falkingham, G. 2005 'A Whiter Shade of Grey: A New Approach to Archaeological Grey Literature Using the XML Version of the TEI Guidelines', *Internet Archaeology* 17 <http://dx.doi.org/10.11141/ia.17.5>
- Fitzpatrick, A. 2012 'Development-led archaeology in the United Kingdom: a view from AD 2010', in R. Bradley, C. Haselgrove, M. Vander Linden and L. Webley (eds) *Development-led Archaeology in North-West Europe*. Oxford: Oxbow, 139-156.
- Fowler, P.J. 1977 *Approaches to Archaeology*. London: A and C Black Ltd.
- Fulford, M. 2011 'The Impact of Commercial Archaeology on the UK Heritage' in J. Curtis, J., M. Fulford, A. Harding, and F. Reynolds, F. (eds) 2011 *History for the Taking? Perspectives on Material Heritage*. London: British Academy, 33-53.
- Fulford, M. and Holbrook, N. 2011 'Assessing the contribution of commercial archaeology to the study of the Roman period in England, 1990-2004', *Antiquaries Journal* 91, 323-345.
- Garwood, P. 2007 'Introduction: the undiscovered country', in P. Garwood (ed.) *The Undiscovered Country. The Earlier Prehistory of the West Midlands*. Oxford: Oxbow, 1-5.
- Gibbs, M. and Holley, S. 2012 'Digital preservation, online access and historical archaeology 'grey literature' from New South Wales, Australia', *Australian Archaeology* 75, 95-103.
- Goacher, H. and Mossop, M. 2011 *Gwel An Mor 2, Portreath, Cornwall, Archaeological Evaluation Interim Report*, Halvasso, Cornwall: Archaeological Consultancy Limited. <http://dx.doi.org/10.5284/1011979>
- Good, O. 2014 *Land adjoining Seven Stars Public House, Baldon Lane, Marsh Baldon, Oxon*, Wessex Archaeology, Salisbury: Wessex Archaeology, 104310.03. <http://dx.doi.org/10.5284/1027871>
- GreyNet n.d. Document Types in Grey Literature. Retrieved from <http://www.greynet.org/greysourceindex/documenttypes.html> [accessed 21 November 2014].
- Habu, J. 2004 *Ancient Jomon of Japan*. Cambridge: Cambridge University Press.
- Hardman, C. 2010 'Our grey literature legacy: opportunity of headache?', *The Archaeologist* 76, 40-41.
- Hardman, C. 2011 'Coming out of the shadows: using Digital Object Identifiers for grey literature', *The Archaeologist* 80, 9-10.
- Hardman, C. and Evans, T. 2010 GLADE: Grey Literature – Access Dissemination and Enhancement. The Pilot Assessment Phase Final Report. York: Archaeology Data Service. Retrieved from <http://archaeologydataservice.ac.uk/attach/research/GLADEreportv5.pdf> [accessed 3 September 2014].
- Harlan, M.E. 2010 'Black and White Literature, Grey Scholars', *Archaeologies: Journal of the World Archaeological Congress* 6(2), 270-288.

- Holbrook, N. and Morton, R. 2008 Assessing the Research Potential of Grey Literature in the study of Roman England. Stage 1 report. Cirencester: Cotswold Archaeology. <http://dx.doi.org/10.5284/1000418>
- Hunter Blair, A. 2009 *Irish Street, Whitehaven, Excavation Report*, Newcastle upon Tyne: AOC Archaeology, 20747. <http://dx.doi.org/10.5284/1024248>
- Hutton, J. and Dickens, A. 2010 *Further Excavations at Langtoft, Lincolnshire. The Glebe Land 2007/8*, Cambridge: Cambridge Archaeological Unit, CAU Report No. 918. <http://dx.doi.org/10.5284/1026287>
- Jeffrey, S. 2012 'A new Digital Dark Age? Collaborative web tools, social media and long-term preservation', *World Archaeology* **44**(4), 553-570.
- Jones, S., MacSween, A., Jeffrey, S., Morris, R. and Heyworth, M. 2001. *From the Ground Up. The publication of archaeological projects: a user needs survey*. York: Council for British Archaeology. Retrieved from <http://www.archaeologyuk.org/publications/puns/> [accessed 2 September 2014].
- Justeson, J.S. 1973 'Limitations of Archaeological Inference: an information-Theoretic Approach with Applications in Methodology', *American Antiquity* **38**(2), 131-149.
- Kansa, E., Kansa, S.W., Burton, M.M. and Standowski, C. 2010 'Googling the Grey: Open Data, Web Services, and Semantics', *Archaeologies: Journal of the World Archaeological Congress* **6**(2), 301-326. <http://dx.doi.org/10.1007/s11759-010-9146-4>
- Landward Research Ltd 2014 *Profiling the Profession* [data-set], York: Archaeology Data Service [distributor]. <http://dx.doi.org/10.5284/1024571>
- Last, J. 2012 'Frameworks for development-led archaeology in England' in R. Bradley, C. Haselgrove, M. Vander Linden and L. Webley (eds) *Development-led Archaeology in North-West Europe*. Oxford: Oxbow, 128-138.
- Lock, G. 2008 'A Professional Mockery', *British Archaeology* July/August, 36-37.
- Louwe Kooijmans, L.P. 2009 'Editorial', *Journal of Archaeology in the Low Countries* **1**, 1.
- Mackenzie Owen, J.S. 1997 'The Expanding Horizon of Grey Literature', *GL3: Third International Conference on Grey Literature, Luxembourg (Luxembourg), 13-14 November 1997*. Retrieved from <http://hdl.handle.net/10760/5654> [accessed 30 November 2014].
- MacLean, S. 2014 HER Content and Computing Survey 2014 Report. Retrieved from https://www.jiscmail.ac.uk/cgi-bin/filearea.cgi?LMGT1=HERFORUM&f=/Surveys/Content_and_Computing/Reports [accessed 06 May 2015]
- Medieval Pottery Research Group 2011 *Medieval Pottery Research Group Bibliography* [data-set], York: Archaeology Data Service [distributor]. <http://dx.doi.org/10.5284/1000154>
- Miles, D. 2013 English Heritage Strategy for Developing Research Resources. Retrieved from <http://www.english-heritage.org.uk/content/imported-docs/a-e/eh-strategy-developing-research-resources.pdf> [accessed 28 September 2014].
- Moore, R.H. and Evans, T.N.L. 2013 'Preserving the Grey Literature Explosion: PDF/A and the Digital Archive' *Information Standards Quarterly* **25**(3), 20-27. <http://dx.doi.org/10.3789/isqv25no3.2013.04>
- Morris, M. 2012 *An Archaeological Excavation in advance of Swinford Windfarm, near Lutterworth, Leicestershire*, Leicester: University of Leicester Archaeological Services, ULAS report 2011-128. <http://dx.doi.org/10.5284/1022233>
- Morrison, W., Thomas, R. and Gosden, C. 2014 'Laying Bare the Landscape: commercial archaeology and the potential of digital spatial data', *Internet Archaeology* 36 <http://dx.doi.org/10.11141/ia.36.9>
- Museum of London Archaeology 2012 *Identification and quantification of archaeological projects arising from aggregate extraction in Hampshire, Surrey, East Sussex and West Sussex* [data-set], York: Archaeology Data Service [distributor]. <http://dx.doi.org/10.5284/1012487>
- Newman, M. 2010 'Curation and conceptualisation: archaeological theory and curatorial archaeology, strange bed fellows?', *Assemblage* **10** [http://www.assemblage.group.shef.ac.uk/pdfs/issue10supp/Newman - Curation and Conceptualisation.pdf](http://www.assemblage.group.shef.ac.uk/pdfs/issue10supp/Newman_-_Curation_and_Conceptualisation.pdf) [accessed 1/10/2014].
- Newman, R., Cranstone, D. and Howard-Davis, C. 2001 *The Historical Archaeology of Britain, c. 1540-1900*. Stroud: Sutton Publishing.
- North Yorkshire HER 2011 A comparative analysis of data from the North Yorkshire HER, Archaeological Investigations Project and OASIS, 2003-2008. North Yorkshire HER: Harrogate.
- OAPEN-UK 2014 Survey of use of monographs by academics. Retrieved from <http://oapen-uk.jiscebooks.org/files/2012/02/OAPEN-UK-researcher-survey-final.pdf> [accessed 3 December 2014].
- OASIS 2014 Results of the OASIS redevelopment surveys start to become available. Retrieved from <http://archaeologydataservice.ac.uk/blog/oasis/?p=129> [accessed 5 December 2014].
- Oikarinen, T. 2014 'Archaeological Grey Reports – Current Issues and Their Potential for the Future' in J. Ikäheimo, A.Salmi and T. Äikäs (eds.): *Sounds Like Theory*. XII Nordic Theoretical Archaeology Group Meeting in Oulu 25-28.4.2012. Monographs of the Archaeological Society of Finland **2**, 187-197.
- Olivier, A. 1996 *Frameworks for Our Past: a Review of Research Frameworks, Strategies and Perceptions*. London: English Heritage.
- Pejsova, P. and Vaska, M. 2011 'An Analysis of Current Grey Literature Document Typology', *The Grey Journal* **7**(2), 73-80.
- Powlesland, D. 2011 'Identifying the unimaginable – Managing the unmanageable', in Cowley, D. (ed.) *Remote Sensing for Archaeological Heritage Management*: Brussels: EAC Occasional Paper No. 5, 17-32.

- Rescue 2013. The fifth report on local government staff resources. English Heritage, The Association of Local Government Archaeological Officers and the Institute of Historic Building Conservation July 2013. An appraisal by RESCUE. The British Archaeological Trust. Retrieved from: <http://rescue-archaeology.org.uk/2013/09/22/heritage-in-crisis-rescue-responds-to-the-fifth-report-on-local-government-heritage-staff-resources/> [accessed 01/12/2014].
- Richards, J.D and Hardman, C. 2008 'Stepping Back from the Trench Edge: An Archaeological Perspective on the Development of Standards for Recording and Publication' in M. Greengrass and L. Hughes (eds) *The Virtual Representation of the Past*, London: AHRC ICT Methods Network.
- Rogers, T. 2010 *Archaeological excavation at the former petrol storage facility, Bath Road, Worcester*, WHEAS, Worcester: Worcestershire Archaeology, Project 2916, Report 1629, WCM 101457. <http://dx.doi.org/10.5284/1011614>
- Roskams, S. and Whyman, M. 2007 'Categorising the Past: Lessons from the Archaeological Resource Assessment for Yorkshire', *Internet Archaeology* 23. Retrieved from <http://intarch.ac.uk/journal/issue23/2/index.html> [accessed 3 December 2014]
- Roth, B.J. 2010 'An Academic Perspective on Grey Literature', *Archaeologies* 6(2), 337-345. <http://dx.doi.org/10.1007/s11759-010-9141-9>
- Rothstein, H. R. and Hopewell, S. 2009 'Grey literature' in H. Cooper (ed.) *The Handbook of Research Synthesis and Meta-analysis (2nd ed)*. New York: Russell Sage, 103-125.
- Schöpfel, J. 2011 'Towards a Prague Definition of Grey Literature', *The Grey Journal* 7(1), 5-18.
- Schöpfel, J. and Farace, D.J. 2010 'Grey Literature' in M. J. Bates & M. N. Maack (eds.), *Encyclopedia of Library and Information Sciences, Third Edition*, CRC Press: Taylor & Francis, 2029-2039.
- Seymour, D. 2010a 'In the Trenches Around the Ivory Tower: Introduction to Black-and-White Issues About the Grey Literature', *Archaeologies: Journal of the World Archaeological Congress* 6(2), 226-232.
- Seymour, D. 2010b 'Sanctioned Inequity and Accessibility Issues in the Grey Literature in the United States', *Archaeologies: Journal of the World Archaeological Congress* 6(2), 233-269.
- Silva, R.T. 2010 'Some Aspects of "Grey Literature" in Brazil', *Archaeologies: Journal of the World Archaeological Congress* 6(2), 327-336. <http://dx.doi.org/10.1007/s11759-010-9142-8>
- Smith, A., Cairns, S. and Vokes, C. 2012 Review of the Development and Implementation of OASIS in England. Pye-Tait Consulting. Retrieved from <http://oasis.ac.uk/pages/attach/PROJECT%20HISTORY/OASIS-Review-Final-Report-270112-v2.pdf> [accessed 12 December 2014].
- Smyth, J. 2014 *Settlement in the Irish Neolithic: New discoveries at the edge of Europe*. Oxford: Oxbow.
- Sommers, M. 2013 *Archaeological Excavation Report: Felixstowe South Seafront and Martello Tower P*, Ipswich: Suffolk County Council Archaeological Service, SCCAS Report No. 2013/083. <http://dx.doi.org/10.5284/1021316>
- Study Group for Roman Pottery 2014 *Study Group for Roman Pottery Bibliography* [data-set], York: Archaeology Data Service [distributor]. <http://dx.doi.org/10.5284/1025061>
- Thomas, R.M. 1991 'Drowning in data? Publication and rescue archaeology in the 1990s', *Antiquity* 65, 822-8.
- Tillett, S. and Newbold, E. 2006 'Grey literature at The British Library: revealing a hidden resource', *Interlending and Document Supply* 34(2), 70-73.
- Tudhope, D., Binding, C., Jeffrey, S., May, K. and Vlachidis, A. 2011 'A STELLAR role for knowledge organization systems in digital archaeology', *Bulletin of the American Society for Information Science and Technology* 37(4). <http://dx.doi.org/10.1002/bult.2011.1720370405>
- Vander Linden, M. and Webley, L. 2012 'Introduction: development-led archaeology in North West Europe. Frameworks, practices and outcomes', in R. Bradley, C. Haselgrove, M. Vander Linden and L. Webley (eds) *Development-led Archaeology in North-West Europe*. Oxford: Oxbow, 1-8.
- Vince, A. 1996 'Editorial', *Internet Archaeology* 1. <http://dx.doi.org/10.11141/ia.1.7>
- Vlachidis, A. and Tudhope, D. 2013 'The Semantics of Negation Detection in Archaeological Grey Literature', *Communications in Computer and Information Science* 390, 188-200. http://dx.doi.org/10.1007/978-3-319-03437-9_20
- Vlachidis A., Binding C., May K. and Tudhope D. 2013 'Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature', in Przepiórkowski, A., Piasecki, M., Jassem, K. and Fuglewicz, P. (eds) *Computational Linguistics – Studies in Computational Intelligence* 458, 187-202. Berlin Heidelberg: Springer.
- Wessex Archaeology 2004 *Addington Street, Waterloo, London: excavation assessment report*, London: Wessex Archaeology, Report ADI04 (55350.01). <http://dx.doi.org/10.5284/1027931>

Library, Information Science & Technology AbstractsTM with Full Text

Available via EBSCOhost[®]

The definitive professional information
resource designed for librarians and
information specialists...

*Library, Information Science & Technology AbstractsTM
with Full Text* is an indispensable tool for librarians
looking to stay current in this rapidly evolving field.

Comprehensive content includes:

- Full text for more than 270 journals and nearly 20 monographs
- Indexing for more than 550 core journals, 50 priority journals and nearly 125 selective journals
- Includes books, research reports, proceedings and author profiles
- Access to 6,800 terms from reference thesauri
- Coverage extends back as far as the mid-1960s

Subject coverage includes:

- Bibliometrics
- Cataloging
- Classification
- Information Management
- Librarianship
- Online Information Retrieval
- And much more...

Contact EBSCO Publishing to learn more about
Library, Information Science & Technology AbstractsTM with Full Text,
or to request a free trial.

Phone: 800.653.2726

Email: request@ebscohost.com

www.ebscohost.com



The recent Italian regulations about the open-access availability of publicly funded research publications, and the documentation landscape in astrophysics *

Monica Marra (*Italy*)

Abstract

In October 2013 Italy enacted Law n.112/2013, containing the first national regulations about the open-access availability of publicly-funded research results (publications). The impact of these new regulations with the specific situation of that open-access pioneering discipline which is astrophysics, has been considered. Under a strictly technical point of view, in the light of the new dispositions the open nature of a part of the astrophysical scholarly literature which has been made available online free to the reader during the last twenty years, might be questionable. In astrophysics, most of the journal articles are published by a very small number of scientific societies and organizations. The copyright policies of these major publishing bodies have been collected and analyzed, with regard to the main requirements of the Italian law about open access. Most of these policies are sufficiently liberal for an entirely compliant open access to be provided and scientists would benefit from knowing these details more extensively. Some possible ways to make astrophysicists' scholarly dissemination entirely compliant with the Italian law requirements are considered.

On October 7th, 2013 Italy achieved an important attainment enacting Law n.112/2013, which is the first national law concerning open access.

The law – which in fact pertains to a number of issues related to cultural heritage - had initially been approved in a different formulation (Decreto-Legge n.91 dated August 8th, 2013). During the passage through Parliament, the first stage of the action was subjected to changes, which according to authoritative commentators (De Robbio 2013; Caso 2013) partly deflected the text about open access from the European Union Recommendation on the subject, issued in July 2012.

Law 112, under article 4, decrees:

- OA is mandatory for research outputs/publications derived from research projects which have been financed by at least 50% with public funds;
- if the publication venue publishes at least 2 issues/year (i.e., according to Caso 2013: books are excluded).
- It may be provided a) either on the publisher's site or b) on an institutional or disciplinary repository, by depositing the publication's accepted version within 18 months (STM area) or within 24 months (SSH area) from publication.
- It must be provided at no additional cost.

During the period that shortly preceded and followed the enactment of Law 112/2013, a series of important initiatives seemed to turn the Italian experts' and concerned people's attention back from the new regulations to the European Recommendation. These initiatives included a) the commitment, signed by the then-Minister for University and Research Massimo Bray, aimed at harmonizing the new law with the European Recommendation (October 3rd, 2013) and consequent actions (i.e., the programme Scientific Independence of young Researchers (SIR) 2014, with open access mandatory clause in the same terms as in Horizon 2020, mentioned below); b) the approval of the 70-billion-euros research financing programme Horizon 2020 at the European Commission, in Autumn-Winter 2013, with European-type open-access clauses; c) the publication of some relevant national statements and recommendations, with a leading role carried out by national academic coordination bodies, also supported by the Presidents of some important national research Institutions, including INAF [1]. These recommendations and statements seem to lean towards the green road to open access and encourage academic and research institutions to issue OA policies and mandates for their research personnel.

* First published in the GL16 Conference Proceedings, March 2015.

In a general perspective, it may be useful to recall that astrophysics is peculiar among other disciplines, with regard to its approach to the dissemination of research outputs. Astrophysics has actually been so deeply crossed by “the preprint culture” during the last twenty years, that in the opinion of most scientists at international level an open-access policy is considered to be already practised (Harley 2010), the proof being generally indicated in the very large adoption of the practice of posting papers on the ArXiv server in a pre-publication stage.

The literature is not unanimous about the percentage of published astrophysical papers already available in open access mode - which is probably related to the different self-archiving practices adopted by researchers according, typically, to different astrophysical subfields. Actually, archiving of the preprint version of papers occurs as well as archiving of the accepted version does; the (limited) posting of post-print versions as well makes this lively scenario even more rich and motley.

Anyway, although it seems unquestionable that astrophysicists' perception of a widely open-mode available disciplinary literature is well-grounded, we might profitably ask ourselves whether researchers on one side, and regulations about open access on the other, are actually considering the very same object.

In fact the Italian law - as well as other national laws concerning open access (Caso 2013) - considers texts as endpoints of the respective scholarly pathways rather than possible stages of research elaboration and dissemination. Thus, regulations require making free to the reader *the papers' final accepted versions* within a given period - either by publishers on their website, or by authors through systematic deposit on a repository. As unexpected as it may seem, a significant fraction of the impressively long “green OA” road which has spontaneously been taken by astrophysicists during the last twenty years might thus result to be non-compliant with the new law requirements, if considered from a strictly technical point of view.

One of the foremost questions for those technically concerned with open access in general is the following: for open access to be provided on authors' side, do publishers let authors manage their manuscripts freely enough? In particular, and key to our perspective: what do publishers allow authors to do with the final accepted versions of their papers?

The crux is authors' rights as stipulated in copyright agreements with publishers, therefore it is essential to be aware of the main journals' policies with regard to this issue.

In astrophysics, “more than 90% of the original research” is published by a very limited number of renowned international journals (Bertout 2012), almost entirely included in the first and second quartile of Thomson Reuters' Journal Impact Factor (JIF) ranking for Astronomy and astrophysics. I have analysed this set of major journals, which includes ApJ (JIF for 2013: 7/59), ApJL (10/59), ApJSS (2/59), AJ (15/59), MNRAS (11/59), A&A (13/59), PASP (17/59).

[1] On March 21st, 2013 the Conference of Rectors of the Italian Universities (CRUI), together with the Presidents of some important Italian research Institutions, signed the *Position statement on open access to the results to research outputs in Italy* (available in English at http://www.cnr.it/sitocnr/lservizi/Biblioteche/Position_statement_OA_en.pdf , last visited Nov. 10th, 2014). This statement is envisioned in a European perspective and commits the signers to coordinating initiatives aimed at i) promoting the creation of open repositories, that be interoperable according to international standards; ii) “encouraging researchers to make their research results (data and publications) available in OA journals or Institutional or subject repositories. Research results deposited in Open Access repositories should be made available in their post-print or publisher's version upon publication, and no longer than 12 months after their publication;” [italics mine]; iii) adopting institutional policies for making deposit mandatory. In January 2014, CRUI issued the Declaration *L'“accesso aperto” alle pubblicazioni scientifiche* (signed jointly with the Italian National University Council (CUN)). The Declaration is critical about the maximum embargo periods stated by the new law and encourages the choice of the “green OA” option (<https://www.cun.it/provvedimenti/sessione/140/dichiarazione/dichiarazione-del-15-01-2014> - in Italian, last visited Nov. 10th, 2014).

All of the journals in this set have both a publisher, and a learned society or scientific organisation which is in fact the reference point for the journals' scientific policy and, which is most important from the present point of view, owns the copyright for the journals' publications and may have a copyright policy of its own, different from the publisher's one. The learned societies or organisations involved are the American Astronomical Society (published by the Institute of Physics Publishing), Royal Astronomical Society (published by Oxford University

Press), European Southern Observatory (EDP Sciences), Astronomical Society of the Pacific (University of Chicago Press).

Thanks to the precious collaboration of these learned societies and organisations, the major journals' policies have been individually checked and a comprehensive view of the options which authors are given was gained - as at Spring 2014.

About the so-called "gold road" to open access – which is in fact notoriously questioned among experts as a recommendable choice for institutions -, three versions turned out to be available.

ApJ, ApJL, ApJSS, AJ, PASP and A&A offered a 12-months delayed open access to the entire content of these serials, free to the reader. The time issue here might arise as a critical element in case regulations or mandates (not to mention the authors' wish) asked for a shorter period for the openness to be provided on the publishers' side. The article processing charge (APC) option was not provided – except for A&A -, but the AAS "welcomes discussions" with interested institutions.

The European journals practised a threefold policy. On one hand, A&A and MNRAS provided niches of their content free to the reader upon publication, while making the remaining full-texts openly available after a period which ranged from twelve months (A&A) to the impressive threshold of three years (MNRAS).

On the other hand, they provided research institutions with the APC option, which corresponds to 400 euros per article for A&A and more than five times as much for MNRAS (as at November 2013).

Obviously, an expenditure issue may arise in some of these cases.

For as much as the "green road" is concerned, ApJ, ApJL, ApJSS, AJ and PASP let authors post the accepted versions of their articles freely on repositories, "whenever they like" for the AAS and also immediately - provided there is an institutional mandate requiring deposit in fewer than twelve months - in the case of PASP.

"Astronomy and Astrophysics" let authors self-archive the accepted version of their papers before publication, but asked researchers to substitute this version with the published one, once the latter is available.

MNRAS let authors self-archive exclusively a PDF version of the final published article, upon publication (as at November 2013).

Bearing in mind all of these copyright policies and the law requirements; considering the citation advantage for papers disseminated in a pre-publication stage; and being notorious that astrophysicists' scholarly routine is in most cases already inclusive of a timely, pre-publication self-archiving of papers on a repository; we might conclude by endorsing the Green Road to open access as a first choice option, thus supporting the recommendation issued by the Italian Conference of Rectors (CRUI), which was signed by the Italian National Institute for Astrophysics on October 4th, 2013 [2].

[2] CRUI and some relevant Italian public research Institutions recommend " [...] the adoption of policies and institutional regulations which make deposit in their institutional repositories mandatory; in case the latter don't exist, publications and data should be deposited in other institutions' repositories or in disciplinary OA repositories".

In case the economics of these processes are a matter of concern, the peculiar modes of dissemination practised by astrophysicists, which have hitherto been characterized by the "productive coexistence" (Henneken et al, 2007) of established journals on one side, and very wide networks of versioned, free online availability on the other side, might preserve the sustainability of the learned societies' activity, or in some cases (MNRAS) hopefully make it more generally affordable to research institutions.

A final, additional consideration deserves attention. In spite of their central importance within astrophysics, the refereed journal articles seem in fact to represent a minority share of the discipline's research output (less than 40% according to the sample study by Marra, 2013). In the spirit of making open access progressively more common practice, the use of a Current Research Information System (CRIS) as an official tool for compliance with the OA provisions might open the way to a similar treatment for the other research and technology output types which are or might be hosted by this kind of tool, and which cooperate significantly in the shaping of the astrophysical documentation landscape.

References

Bertout, Charles (2012), *A brief overview of the publication process in astronomy*, in *id.*, *A guide to effective publishing in astronomy*, Les Ulis, EDP Sciences, 2012, 13-18.

Caso, Roberto (2013), *La legge italiana sull'accesso aperto agli articoli scientifici: l'inizio di un percorso normativo*, in ROARS. Return On Academic ReSearch (blog), 6.11.2013, <http://www.roars.it/online/la-legge-italiana-sullaccesso-aperto-agli-articoli-scientifici-inizio-di-unpercorso-normativo/>

De Robbio, Antonella (2013) *L'Open Access è legge anche in Italia*, in *Il BO. Il giornale dell'Università degli Studi di Padova*, 25.10.2013, <http://www.unipd.it/ilbo/content/l%E2%80%99open-access-e-legge-anche-italia>

Harley, Diane et al. (2010), *Astrophysics case study*, in *id.*, *Assessing the future landscape of scholarly communication: an exploration of faculty values and needs in seven disciplines*, Berkeley, Center for Studies in Higher Education, 2010, 138-206.

Henneken, Edwin et al. (2007), *E-prints and journal articles in astronomy: a productive coexistence*, "Learned Publishing" 20(1), 2007, 16-22.

Marra, Monica (2013), *Astrophysics as an excellent object for bibliometric analysis, myth or reality? A sample study comparing bibliographic coverage in NASA/ADS, Scopus, Web of Science and in the institutional database CRIS-INAF* (not published), DOI: 10.13140/2.1.1493.8561 https://www.researchgate.net/publication/261911739_Astrophysics_as_an_excellent_object_for_bibliometric_analysis_myth_or_reality_A_sample_study_comparing_bibliographic_coverage_in_NASAADS_Scopus_Web_of_Science_and_in_the_institutional_database_CRIS-INAF?ev=prf_pub

Marine Planning and Service Platform (MAPS) An Advanced Research Engine for Grey Literature in Marine Science *

Sara Goggi, Monica Monachini, Francesca Frontini, Roberto Bartolini,
Gabriella Pardelli, Maurizio De Mattei, Franco Bustaffa, and
Giuseppe Manzella (Italy)

Abstract

The MAPS (Marine Planning and Service Platform) project is a development of the **Marine** project (Ricerca Industriale e Sviluppo Sperimentale Regione Liguria 2007-2013) aiming at building a computer platform for supporting a Marine Information and Knowledge System, as part of the data management activities. One of the main objective of the project is to develop a repository that should gather, classify and structure marine scientific literature and data thus guaranteeing their accessibility to researchers and institutions by means of standard protocols.

We will present the scenario of the Operative Oceanography together with the technologies used to develop an advanced search engine which aims at providing rapid and efficient access to a Digital Library of oceanographic data. The case-study is also highlighting how the retrieval of grey literature from this specific marine community could be reproduced for similar communities as well, thus revealing the great impact that the processing, re-use as well as application of grey data have on societal needs/problems and their answers.

1. Introduction

Operational Oceanography is the branch of marine research which deals with the development of integrated systems for monitoring, analyzing, modeling and forecasting oceans and seas. The integrated systems need access to real-time as well as delayed mode data. The operational oceanography activities includes also the re-analysis of data and models to assess marine variability or long term trends. These activities require the access to quality controlled data and to information that is provided in reports and/or in relevant scientific literature. This finds application in many areas, ranging from environmental studies, security and safety to protection of off-shore and coastal infrastructures.

Hence, creation of new technology is needed by integrating several disciplines such as data management, information systems, knowledge management and others. Full-text technologies are often unsuccessful when applied to this type of queries since they assume the presence of specific keywords in the text; in order to fix this problem, the MAPS project suggests to use different semantic technologies for retrieving the text and data and thus getting much more complying results.

In this scenario the search engine (which is under development) has been designed for operating with structured data – the metadata – and raw data – the associated technical and scientific documentation. The set of metadata associated with marine data is defined in the CDI (Common Data Index) documented standard developed in the framework of the EC project SeaSearch/SeaDataNet. They encode: the types of sizes which have been measured; the measurement tools; the platform which has been employed; the geographic area where measures have been taken; the environmental matrix; the descriptive documentation.

This paper is organised as follows: Section 2 is presenting the related work; Section 3 the methodology adopted and the search engine to be developed; Section 4 shows a case-study and gives examples for using the MAPS search engine.

2. Related work

"Information retrieval (IR) research has been actively driven by the challenging information overload problem and many successful general-purpose commercial search engines. While the popularity of the largest search engines is a confirmation of the success and utility of IR, the identification, representation, and use of the often-complex semantics behind user queries has not yet been fully explored. In this workshop we target methods that exploit semantics in search-related tasks. One of the major obstacles in bridging the gap between IR and Natural Language Processing (NLP) is how to retain the flexibility and precision of working with text at the lexical level while gaining the greater descriptive precision that NLP provides. We have solicited

* First published in the GL16 Conference Proceedings, March 2015.

contributions on automatic analysis of queries and documents in order to encode and exploit information beyond surface-level keywords: named entities, relations, semantic roles, etc" (Workshop on Semantic Search, NAACL 2010).

The term Text Mining (TM) denotes all those semantic technologies that aim to extract structured information from un-structured data (Ramjan et al. 1998). Information Retrieval (IR) groups instead the technologies aiming at retrieving certain information from a document database (Manning, et al. 2009). The currently popular search systems, including those implemented by common search engines, are mainly based on matching of strings: the words of the query are treated as a set of key terms whose presence or absence is sought in previously indexed texts (e.g. Brin, 1998; Kogut et Holmes, 2001)

Semantic search systems are slowly spreading amongst the generalist search engines. They are used in particular in specific domains contexts where the reduced domain terminology makes the adoption of semantic technologies easier. The leading field of these systems is the biotechnology one where the experts need to quickly access to large amounts of information about the literature domain by using specific research related to concepts and semantic relations (Ananiadou et al. 2010, Dai et al. 2010). Even for biomedicine it is important to being able to extract conceptual information about the interaction between the genes and/or proteins; finally, also in the field of environmental sciences significant experiences in this sense have been made.

In general, a preliminary analysis of the texts in question by means of existing technologies relying on lexical and conceptual generic resources it is always necessary for the realisation of a statistical semantic search system in a specific domain. Later on, the results of this analysis – possibly manually corrected - are used for extracting new knowledge through specific terminology that will enrich the lexical resources: in this way these resources will act as a training set for the linguistic analysis systems.

Semantic search systems, like any other search system, may propose systems for search refinement based on metadata or on information previously extracted from the text. One of the most useful aspects of MAPS is the geographic profiling of documents, both on the basis of metadata and of indication of geographical places in the text. In this sense, also systems for the analysis and extraction of named-entities are relevant to the search.

We present here some relevant projects in which semantic search technologies and/or textual analysis have been developed for the extraction and retrieval of information from domain-specific texts:

OpeNER is a project funded by the European Commission under the 7th Framework Programme. Its acronym stands for "Open Polarity Enhanced Name Entity Recognition". The main objective of the project is to provide a set of tools which can be immediately usable by researchers from small and medium enterprises for performing natural language processing activities which are free and easy to integrate into their workflow. Amongst the purposes of OpeNER is to succeed in detecting and disambiguating named-entities as well as performing sentiment analysis and opinion detection on texts, so as to be able to extract from web reviews the feeling and the customers' opinion about a certain resource.

BOOTStrep - Bootstrapping of Ontologies and Terminologies Strategic Research Project - has been funded by the European Commission 6th Framework Programme. The project operated in the biological domain and combined terminology resources and existing databases in a common framework. BOOTStrepp implemented a system of text analysis for the acquisition of new terms, concepts and relations which is a semi-automatic process of ontology-construction in the biological field.

KYOTO (Knowledge Yielding Ontologies for Transition-based Organization), [Co-funded by EU - FP7 ICT Work Programme o2007 under Challenge 4 - Digital libraries and Content] had the goal of providing a system for semantic search in order to allow expert users to model and improve their domain ontology with terms and concepts automatically extracted. The application domain of KYOTO was the environment.

GLOSS - Global Semantic System - The overall objective of the Italian project GLOSS - derived from the KYOTO project but applied to the geographical domain- was focused on building a global system for the creation, use and sharing of documents starting from the current state of the art of the MLIA2MMC technology. The project produced an integrated environment for consultation and sharing of knowledge.

3. Methodology

After having analysed the state of the art, the development of the MAPS project has followed these steps:

- 1) Collection of needs. This phase focused on the detection of relevant information regarding the content of documents. The information was obtained from interviews and questionnaires to domain users in order to highlight the specifications for designing the engine.
- 2) Modelling of the Digital Library. This phase was based on the identification of the optimal pipeline according to the users' requirements as collected in Step 1 and on the definition of the requirements regarding the lexical/terminological/ontological resources necessary for the semantic search.
- 3) Definition and design of the search engine architecture. This third phase was further divided in: a) Detection of specific semantic search interface in order to determine which are the requirements to be met; b) implementation of an effective and performing instrument for document search with respect to the needs of the oceanographic community.
- 4) Selection of information for the case-study presented in this work.

The design of the semantic engine architecture is based on the integration of already existing linguistic/semantic modules emphasizing the adaptation to the lexical and terminological document bases of the specific marine domain, such as the Italian Marine Semantic Network.

The collaboration between the various partners of the project - ILC-CNR of Pisa and the groups affiliated to the DP2000 and ETTsolutions companies (of La Spezia and Genoa, respectively)- resulted in the elaboration of formats of linguistic and semantic text annotation and of standards for lexical and ontological resources, in order to actually make possible the indexing and querying of documents processed in the database of the Marine project.

3.1. Search Engine

The MAPS Document Base is formed by a set of Italian and English pdf texts and they are the technical and scientific documentation which describes the marine data made available by MAPS. Each text belonging to the catalogue is converted into text format and then processed by a battery of linguistic analysis modules which enrich it with multiple layers of stratified features corresponding to the levels of analysis performed. This expansion is followed by a phase of contraction due to the extraction of semantic content within the text.

With reference to the architecture in Figure 1, the Indexing module collects the extracted content (domain terms, concepts and relationships between them) and it anchors them to the source text in order to facilitate retrieval at a later stage. The Cluster Builder module, instead, organises the extracted semantic information in a domain ontology.

The extraction of information is conducted in several logical steps

- creation of the glossary of terms;
- creation of the network between terms and concepts (ontology taxonomy);
- creation of the profiled text (indexing) belonging to the catalogue of documents.

The terminological glossary consists of a list of automatically extracted terms that – for the English part –will enrich an already existing terminology database: the Oceanographic BODC vocabulary (British Oceanographic Data Center). However, the texts of the catalogue will always be accompanied by the metadata in CDI format (Common Data Index) from which it will be possible to extract some meta-information for creating a small document base for Italian as well.

The name-entities belonging to the glossary may be both simple and complex: the simple terms are those extracted from the NER (Named Entity Recognition) module, that is dates, geographical names, place names, organisations, abbreviations, acronyms, etc. and those terms consisting of a single word, normalised by lemma and filtered by their frequency in documents. The complex terms are noun phrases linguistically normalised which correspond to patterns (such as name-name, name-preposition-name etc.) identified by a sort of mini-grammar. The complex terms thus identified are sorted according to a statistical rank and should be evaluated by applying measures of association (log-likelihood, mutual information, etc.).

The domain ontology is a graph representing a semantic-conceptual network; the network nodes are the terms and concepts of the glossary. The taxonomy is constructed by identifying the hyponymy and hyperonymy relations between terms and this process can be done:

- on a lexical basis - the system is in fact capable of identifying terms with increasing complexity and the sharing of the lexical head is a proof of hyperonymy; for example, if the system identifies as domain terms "boa" and "signal boa" then it determines that "boa" is a hyperonym of "signal boa" and assigns a hyponymy relation in the opposite direction. Similarly, if you find the term "dangerous signal boa" the same reasoning applies to a level further;
- exploiting lexical-semantic resources such as WordNet; in this case, at least for the more generic terms, hyponymy and hyperonymy relations are imported directly from WordNet thus allowing also the extension of the glossary with new concepts.

The introduction of hyperonymy and hyponymy relations concerns the construction of vertical relationships within the ontology. Horizontal relations between terms are partly built by exporting the relationships between entities in Wordnet - if present - but mostly by performing a dependency parsing of texts and identifying the verbs which have a relation with the domain terms.

Finally the indexes allow to efficiently retrieve interesting textual documents when the system is semantically queried on the basis of a specific user's need; for this purpose the indexing process allows to associate each text of the catalogue to a list of terms and concepts that represent it maintaining the reference to the original text.

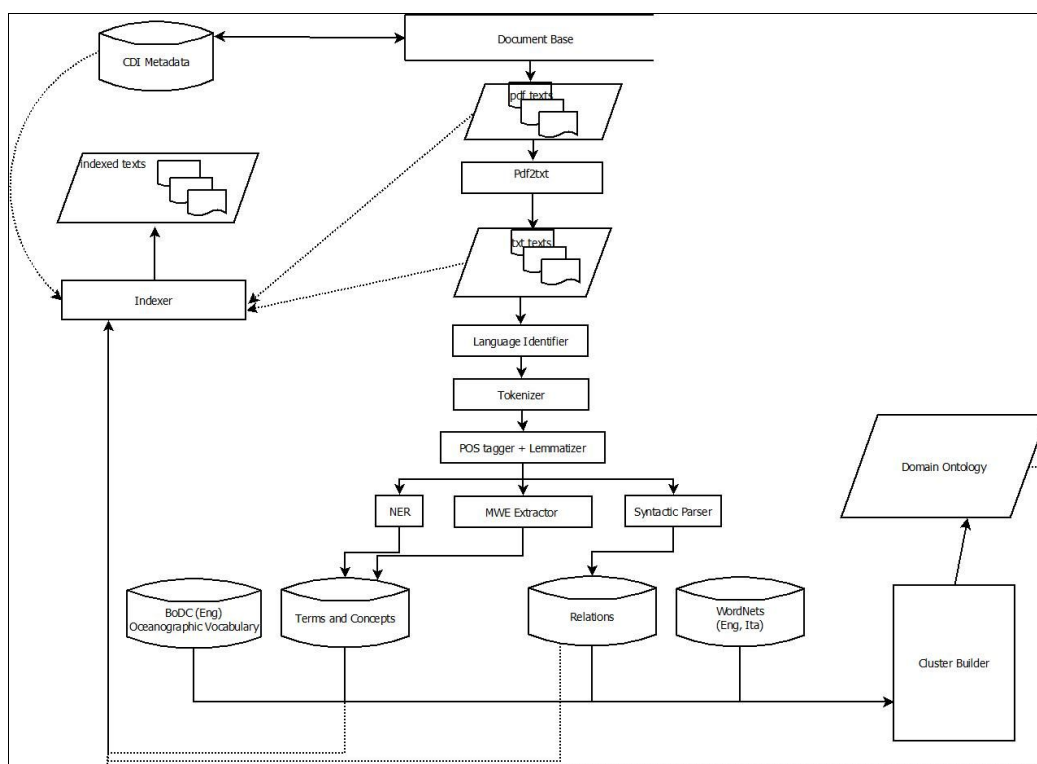


Figure 1: Linguistic Analysis and Indexing

4. A case-study

4.1. Terms

This case-study highlights how the retrieval of grey literature from this specific marine community could be reproduced for other communities as well, thus revealing the great impact that the processing, re-use as well as application of grey data have on societal needs/problems and their answers. The marine application is also making use of common vocabularies that are defined by an international working group called 'SeaVox'.

The illustrative term chosen from the digital library is *Temperature*¹: this noun has been captured by the semantic engine in the wide variety of contexts used in oceanographic scientific literature. The term *Temperature* is often associated with other nouns, adjectives, acronyms and abbreviations thus forming complex terms belonging to the domain ontology; these terms in relation with *Temperature* are normalised (lemmatised) in order to be independent from the

¹ For visualising some associated forms of the term *temperature* or some complex terms related to it, please see *Appendix 1*.

lexical form: singular/plural, lowercase/uppercase, graphical variants (for example, *sea surface temperature*, *sea-surface temperature*).

The forms extracted by the oceanographic repository and processed by the semantic engine reveal that the language of this field is mostly constituted by technical terms as well as by acronyms and abbreviations: they might become significant anchors for the information retrieval and the knowledge extraction and this is the reason why the acronym SST (Sea Surface Temperature)² has been picked up and the attention has been focused on contexts where it occurs.

It can be noted that in the ontology various concepts gather around the domain term SST: “*temperature*”, “*surface temperature*”, “*sea surface temperature*” are all terms linked together by hyponymy and hyperonymy relations while “*sea surface temperature*” e “SST” stand for the same concept and are therefore linked by an identity relation.

Figure 2 below shows a few terminological examples extracted from documents where they appear either as a simple term or as belonging to a complex term; the ontological concept has been underlined and is followed by a list of micro-contexts where it is present.

<u>SST</u>	<u>SST Anomaly</u>
- the night time SST - Mediterranean SST - satellite SST - a 1/16 degree resolution Black Sea SST - a 1/100 degree resolution Black Sea SST - The very low SST - The very high SST	SST anomalies
	<u>SST Retrieval</u>
	SST retrieval
	<u>Information SST Level</u>
	practical information on the SST TAC level
<u>SST measurement</u>	<u>SST Analysis</u>
- in-situ SST measurements - valid SST measurements - input satellite SST measurements - buoy SST measurements	- the OSTIA³ SST analysis - SST analysis - the global SST analysis
<u>SST Measurement Area</u>	<u>SST TAC⁴</u>
input satellite SST measurements in the Arctic area	SST TAC
<u>SST Satellite Region</u>	<u>Local SST Characteristic</u>
SST from NOAA⁵ satellite for the Australian R region	local SST characteristics
<u>Real-Time SST</u>	
near real-time from various satellite SST producers	

Figure 2 – A few terminological examples

² *Sea surface temperature*: “Sea surface temperature (SST) is the water temperature close to the ocean's surface. The exact meaning of surface varies according to the measurement method used, but it is between 1 millimetre (0.04 in) and 20 metres (70 ft) below the sea surface” (From Wikipedia).

³ Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA). http://ghrsst-pp.metoffice.com/pages/latest_analysis/ostia.html

⁴ Thematic Assembly Centres

<http://en.wikipedia.org/wiki/MyOcean#TAC> .28Thematic Assembly Centres.29 Production Centres.

“The MyOcean project, which is funded by the European Union for a 3-year period (2009-2012), includes the development and operations of a Thematic Assembly Centre for satellite SST products (SST TAC) from (MYOCEAN RDAC PROGRESS REPORT, 2012)..

⁵ National Oceanic and Atmospheric Administration (NOAA). <http://www.noaa.gov/about-noaa.html>

4.2 Oceanographic Semantic Extraction

4.2.1. Digital Library for Grey Documentation

The aim of oceanographic semantic extraction is to retrieve relevant information from digital documents: the innovation of a semantic engine lies in the fact that the process is not just about the retrieval of already known documents by means of a simple term query but rather the retrieval of a population of documents whose existence was unknown.

a) The system answers by showing a screenshot of results ordered according to the following criteria:

- Relevance Relevance of the document with respect to the “*concept*” which is searched
- Date Date of publication of the paper
- Source Providers who share the scientific document
- Matrix Environmental matrices as defined in the oceanographic field
- Geographic area Relevance of the geographic area specified in the query
- Clustering “The process of organizing objects into groups whose members are similar in some way”⁶.

≈ The clustering returns as the output the related documents, that is those documents which have in common more than one “concept” described in the ontology.

b) For each document the visualization provides:

- Title, author, source/provider, web address
- Tagging of key terms or concepts
- Summary of the document
- Visualisation of the whole document

c) Advanced search and interaction with the results: the possibility to further refine the search for reducing/increasing the number of results by means of the following partition:

- Typology of formats Pdf, doc, xls, ppt, etc.
- Typology of material Video, Audio, etc.
- Language of the document English, Spanish, French, German, etc.
- Authors’ affiliation Academy, Industry, etc.

The possibility of inserting the number of citations for each document among the criteria of the advanced search is currently undergoing; in this case the engine should be able to connect to any of the existing bibliographic citation systems (such as Google Scholar, Scopus, etc.).

The following is a screenshot of the MAPS semantic search engine:

⁶ http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/.

MOTORE DI RICERCA SEMANTICO

MAPS – Digital Library

Inserisci il testo da cercare

Salinità relativa all'aumento della temperatura

Cerca

Esempio: salinità, rilevazione, ...

Numero totale dei documenti trovati

Ontologia trovata

Visualizzazione 1 – 20 di 321 risultati in 0.84 secondi di **salinità relativa aumento temperatura**

Predizione: 0.8

Ordinamento per rilevanza

Forse cercavi: **salinità relativa alla temperatura**

- G. Zodiatis, R. Lardner, G. Georgiou, E. Demirov, N. Pinardi, G. Manzella. The Cyprus coastal ocean forecasting and observing system. **Fonte:** Sea Technology **Anno:** 2003 **Provider:** ENEA **Percorso:** <http://www.enea.it/docs/The Cyprus coastal ocean forecasting and observing system.pdf>
- G. Borzelli, G. Manzella, S. Marullo, R. Santoleri. Observations of coastal filaments in the Adriatic Sea. **Fonte:** Journal of Marine **Anno:** 1999 **Provider:** CNR **Percorso:** <http://www.cnr.it/docs/Observations of coastal filaments in the Adriatic Sea.pdf>
- S. Vignudelli, M. Astraldi, G.P. Gasparini, P. Cipollini, G. Manzella. Integrated use of altimeter and in situ data for understanding the water exchanges between the Tyrrhenian and Ligurian seas. **Fonte:** Journal Geophysical Research **Anno:** 2000 **Provider:** UNIGE **Percorso:** <http://www.unige.it/docs/Integrated use of altimeter.doc>
- S. Sparnocchia, P. Picco, G.M.R. Manzella, A. Ribotti, S. Copello, P. Brasey. Intermediate water formation in the Ligurian Sea. **Fonte:** Oceanologica Acta **Anno:** 1995 **Provider:** UNINA **Percorso:** <http://www.unina.it/docs/Intermediate water formation in the Ligurian Sea.doc>

Documenti correlati:

- G. Borzelli, G. Manzella, S. Marullo, R. Santoleri. Observations of coastal filaments in the Adriatic Sea. **Fonte:** Journal of Marine Systems **Anno:** 1999 **Provider:** ARPAL **Percorso:** <http://www.arpal.it/docs/Observations of coastal filaments in the Adriatic Sea.pdf>

Restrizioni ricerca

Matrici Ambientali

Aria (15391)

Acqua (12527)

Sedimenti (5392)

Bio (7383)

Fonte/Provider

CNR (15391)

ARPAL (12527)

UNIGE (5392)

ENEA (7383)

mostra tutti ...

Anno

2014 (15391)

2013 (12527)

2012 (5392)

2011 (7383)

mostra tutti ...

Gruppi di parametri

Seleziona ...

Area geografica

Seleziona ...

Tipo di documento

doc (12391)

pdf (18527)

ppt (4392)

xls (2338)

Figure 3 – Search example

5. Conclusions

The MAPS project which has been described in this paper aims ultimately at building a computer platform for supporting a Marine Information and Knowledge System. This integrated MAPS platform offers the advantage to have all content in one place and allows linking different information that exists, thus helping data management activities.

It is well-known that document indexing techniques are not sufficient to satisfy user information needs that go beyond the limits of a simple term matching search. Therefore the MAPS search engine is enriched with semantic technologies aimed at providing an accurate representation of the content of vast repositories of unstructured documents for semantic indexing purposes. Today language technologies make it possible for scientists and developers to produce software applications capable of revealing the semantic properties of textual elements and associating them with conceptual structures. Search functions coupled with semantic-conceptual technologies are able to interpret the underlying search criteria and thus better identify the data and the corresponding documents. This enables an effective and selective access to information even in the presence of significant collections of data.

The proposed solution combines the data available with the technical and scientific documentation thus allowing searching data on the basis of the content of these documents. This makes it possible to select data with greater flexibility, using information that cannot be encoded in predefined data structures. Search functions able to capture technical-scientific documentation associated to a query constitute an essential feature for the development of

advanced services for Operational Oceanography, in sectors such as e.g. civil and military security, safety of navigation and protection of offshore and inshore infrastructures.

While semantic technologies have already been used in other areas for information retrieval with excellent results, their use in oceanography is innovative and responds to information needs that only recently emerged: the interest in the marine environment, so far relegated primarily to the scientific world, is in fact investing new sectors under the actions of the European Community and greater sensitivity of the society for the protection of sea life.

This study has attempted to provide semantic annotations to grey literature documents of the oceanography domain. The initial experiment has revealed that available methods are capable of assisting the process of semantic annotation with promising results. The incorporation of ontologies and knowledge resources in a rule-based information extraction technique promises to enable rich semantic indexing of grey literature documents.

Flexibility is a necessary requirement for any new use that the system intends to embrace while additional efforts are required for further exploitation of the technique and adoption of formal evaluation methods for assessing the performance of the method with measurable criteria.

Acknowledgements

This paper is supported by the Programma Operativo Regionale POR-FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 "Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012. Concessione di agevolazione POS. N° 19. CUP G45C13000940007 MAPS.

References

- FRONTINI F., BARTOLINI R., MONACHINI M., PARDELLI G. MAPS. (2014). Stato dell'Arte dei motori semantici . PROGRAMMA OPERATIVO REGIONALE POR-FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 "Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012. CONCESSIONE DI AGEVOLAZIONE POS. N° 19. CUP G45C13000940007 MAPS. Deliverable D1.1, Versione 1.0.
- BUSTAFFA F., DE MATTEI M. (2013). *MAPS. Raccolta delle esigenze*. . PROGRAMMA OPERATIVO REGIONALE POR- FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 "Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012. CONCESSIONE DI AGEVOLAZIONE POS. N° 19. CUP G45C13000940007 MAPS. Deliverable D1.2, Versione 1.0.
- ANANIADOU, S., PYYSALO, S., TSUJII, J. I., & KELL, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7), 381-390.
- DAI, H. J., CHANG, Y. C., TSAI, R. T. H., & HSU, W. L. (2010). New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology*, 25(1), 169-179.
- FENG D., CALLAN J., HOVY E., PASCA M. (Eds.)(2010). *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search* [June 5, 2010, Los Angeles, California]. Association for Computational Linguistics (ACL), Stroudsburg, PA. Page iii.<<http://aclweb.org/anthology/W10-1200>>
- MANNING CHRISTOPHER D., RAGHAVAN P., SCHÜTZE H. (2009). *An Introduction to Information Retrieval*. Online edition, Cambridge UP. <<http://nlp.stanford.edu/IR-book/pdf/00front.pdf>>
- KOGUT P., HOLMES W. (2001). AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. In *First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation*. <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.8180&rep=rep1&type=pdf>>
- BOURIGAULT D., JACQUEMIN C., L'HOMME M.C. (Eds.) (2001). *Recent Advances in Computational Terminology*. John Benjamins Publishing Company, Amsterdam-Philadelphia. Pages viii – xvii.
- OAKES MICHAEL P., PAICE CHRIS D. (2001). Term extraction for automatic abstracting. In Bourigault D., Jacquemin C., L'Homme M.C. (Eds.), in *Recent Advances in Computational Terminology*. John Benjamins Publishing Company, Amsterdam-Philadelphia. Pages 353-370.
- BRIN S. (1998). Extracting Patterns and Relations from the World Wide Web. In *Proceedings of the WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*. Pages 172-183.
- RAJMAN, M., & BESANÇON, R. (1998). Text mining: natural language techniques and text mining applications. In *Data Mining and Reverse Engineering*. Springer US. Pages 50-64.

Enhancement of the functions of the Japan Atomic Energy Agency Library's Fukushima Nuclear Accident Archive using a novel data flagging system that improves the utilization of numerical data on the Internet *

Kiyoshi Ikeda, Mayuki Gonda, Shun Nagaya, Misa Hayakawa, Yukinobu Mineo, Katsuhiko Kuni, Minoru Yonezawa, and Keizo Itabashi (Japan)

Abstract

Related to the Fukushima Daiichi Nuclear Power Station Accident, the Japan Atomic Energy Agency Library has accumulated valuable information on Internet. In the Fifteenth International Conference on Grey Literature, we reported the development of the Fukushima Nuclear Accident Archive using the DSpace. We have encountered a new, challenging issue of grey literature. In many cases, Internet information contains valuable numerical data. However, identifying the existence of numerical data on Internet sites is difficult, and the metadata created for the Fukushima Nuclear Accident Archive cannot currently be used to distinguish whether such information contains numeric data. Therefore, we have considered a method to identify numerical data and have introduced a "data flagging" system that has been used in the International Atomic Energy Agency's International Nuclear Information System. In this paper, we introduce the proposed data flagging system and discuss its application to the Fukushima Nuclear Accident Archive.

1. Introduction

The Fukushima Daiichi Nuclear Power Station Accident (hereafter referred to as "the Fukushima Accident") occurred in March 2011. Information about the accident has primarily been published via the Internet, and permanent accessibility to the information, bibliographic control, and so on have been critical issues in the collection, preservation, and provision of Internet information related to the Fukushima Accident. We previously reported the development of an archive using DSpace to solve such issues and have discussed "Internet information is Grey Literature" as a new type of grey literature at the Fifteenth International Conference on Grey Literature [1].

A widely diverse and massive amount of data is utilized on the Internet. In many cases, Internet information contains valuable numerical data, such as observed, compiled, evaluated, experimental, statistical, and theoretical data.

However, determining the existence of numerical data on Internet sites and distinguishing whether information contains numeric data using our created metadata for the archive is difficult. An effective method to identify whether information contains such data would be very convenient.

We have considered a method to identify numerical data and attempted to introduce a "data flagging" system, which has been used in the International Atomic Energy Agency's International Nuclear Information System (INIS). In the INIS, data flagging is used to identify the existence of useful data in documents and literature.

In this paper, we introduce our data flagging system for numerical data on the Internet and its application to the Fukushima Nuclear Accident Archive.

2. Status of the Fukushima Nuclear Accident Archive

2.1 Collection of Fukushima Accident Information

The Japan Atomic Energy Agency (JAEA) Library is one of the largest nuclear information centers in Japan. After the Fukushima Accident, we have supported scientific research and development activities related to the Fukushima Accident, and the archive is one of the major missions of the JAEA Library [2][3]. Since April 2011, we have collected, compiled, and distributed Fukushima Accident's reference information using a special website [Fig. 1].

* First published in the GL16 Conference Proceedings, March 2015.

This website consists of

- (1) research and development results related to the Fukushima Accident by JAEA staff,
- (2) bibliographies related to the Three Mile Island Unit 2 reactor, Chernobylsk-4 reactor accident, environmental decontamination, and other topics (19 themes, approximately 18,000 articles), and
- (3) Internet sources related to the Fukushima Accident (30 fields, approximately 4,000 URLs).



Fig. 1 JAEA Library website related to the Fukushima Accident

Another role of the JAEA Library is to act as an archive for the INIS National Center of Japan. The INIS is an open access database of published scientific literature on the peaceful uses of nuclear science and technology. The JAEA Library has provided database input records for literature (books, articles, conference proceedings, and technical reports) related to the Fukushima Accident to the INIS.

However, much of the information related to the Fukushima Accident, such as records, documents (press releases and reports), photos, videos, and numerical data (e.g., radiation monitoring and plant data), is published via the Internet, which makes an entire coverage difficult. Parts of this Internet information has been distributed as grey literature because of a lack of bibliographical control and some websites not having permanent accessibility.

Thus, in 2013, we began to collect, organize, and disseminate such information because preserving the observations and knowledge obtained from the Fukushima Accident is very important.

2.2 Outline of Fukushima Nuclear Accident Archive

The Fukushima Nuclear Accident Archive is a bibliographic database that covers Internet information of Japanese government-based organizations and the Tokyo Electric Power Company (TEPCO), as well as bibliographic information about oral presentations regarding credible nuclear associations or societies related to the Fukushima Accident [Fig. 2]. The Fukushima Nuclear Accident Archive was opened to the public on June 23, 2014, to support related research and development [Fig. 3] [4][5].

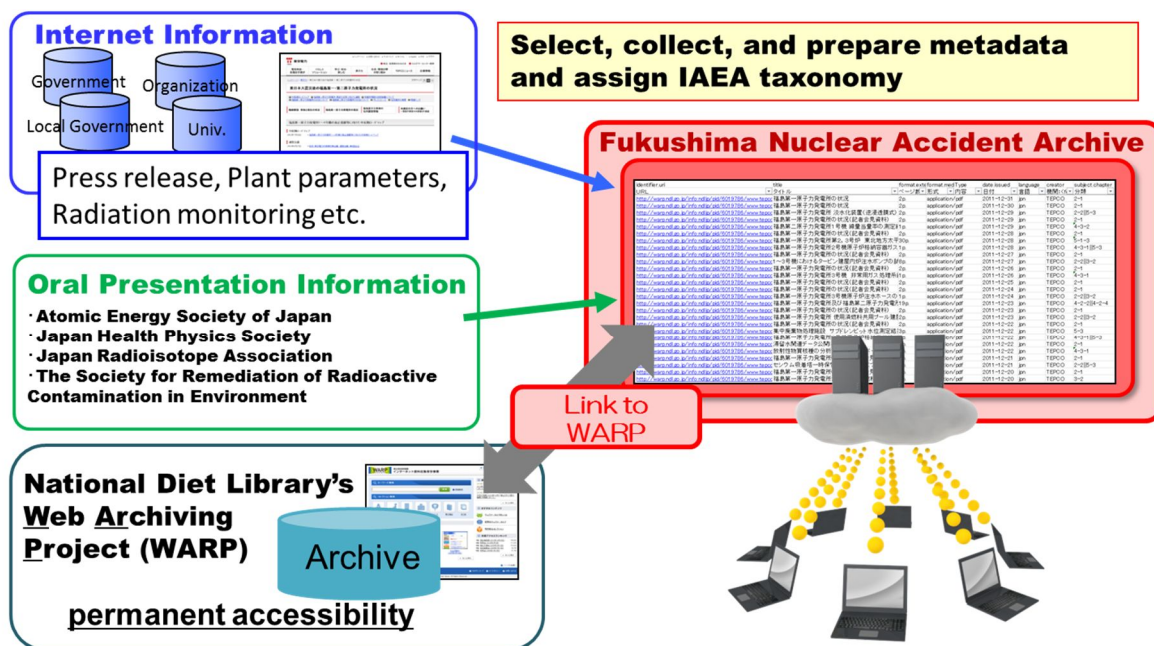


Fig. 2 Outline of the Fukushima Nuclear Accident Archive

To ensure permanent access to this Internet information, we adopted the National Diet Library's Web Archiving Project (WARP) [6]. The WARP stores official Japanese institutions' websites, such as those related to the federal government, the Diet, the courts, local governments, independent administrative organizations, and TEPCO.

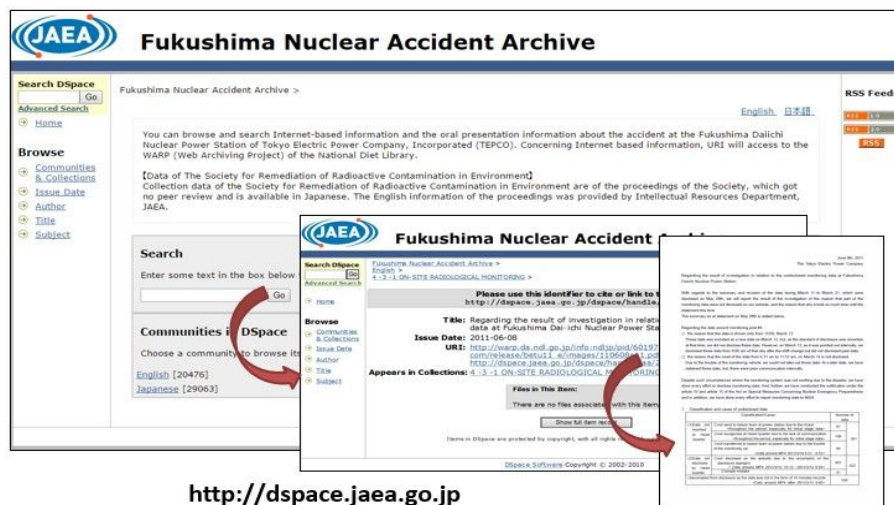


Fig. 3 Fukushima Nuclear Accident Archive

In the archive, we employ DSpace as the archive system because it is commonly used as an institutional repository system. DSpace supports the Open Archives Initiative Protocol for Metadata Harvesting (hereafter referred to as “OAI-PMH”) and is capable of processing standardized metadata.

In addition, we also employ the Dublin Core Metadata Elemental Set (hereafter referred to as “Simple Dublin Core”) and use the IAEA’s Nuclear Accident Knowledge Taxonomy as a classification system.

We provide more than 50,000 metadata for Internet information related to the Fukushima Accident, such as records, documents, photos, videos, and numerical data, from agencies such as TEPCO; the Ministry of Economy, Trade and Industry; the Ministry of Education, Culture, Sports, Science and Technology; and the Ministry of the Environment.

3. Data Flagging System to Utilize Internet Information with Numerical Data

3.1 Numerical Data

Tremendous amounts of numerical data exist that have been used in various fields of scientific research and society. Numerical data refer to physical quantities, such as densities, melting points, cross sections, and spectral shifts, or the number of occurrences of certain events, such as failures of a piece of equipment and coefficients or parameters (such as the rate constant of a particular chemical reaction).

In the nuclear field, Internet information contains valuable and useful data, e.g., radiation monitoring data, plant parameters, physical quantities, and processed data, such as compiled, evaluated, experimental, statistical, and theoretical data.

Numerical data on the Internet has three issues.

- (4) Identifying its existence is difficult.
- (5) The structural aspects of the data, such as file format, content type, and whether it is raw or processed data, must be determined.
- (6) The validity and reliability of numerical data acquisition methods require evaluation.

In many cases, no information exists in the metadata that indicates whether the data contains numerical data. In addition, information often exists in a mixed state of various types of documents (photos, videos, text, and so on) [Fig. 4]. Thus, the metadata must identify whether such data contains numerical data.

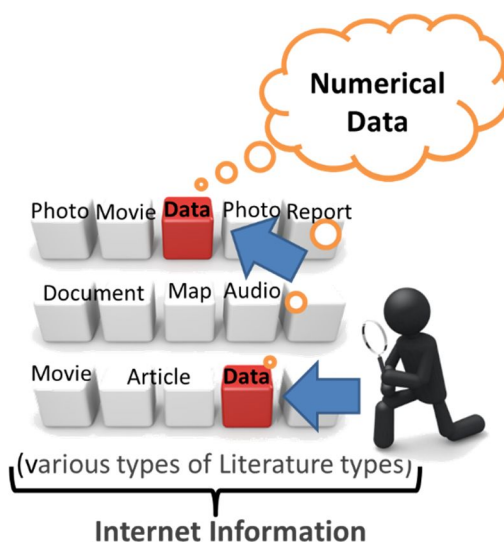


Fig. 4 Various types of information

Another issue is related to the reuse of information. For example, numerical data in PDF or JPEG format is inconvenient; it must be converted to comma-separated value (CSV) format by processing for effective use. In addition, the quality of the information is often an issue. The latter two issues can only be addressed after the information has been obtained. In this paper, we introduce efforts to solve the first issue.

For many people engaging in research and development, availability of a useful method to identify whether information contains useful data would be very advantageous. Thus, we introduce a data flagging system to identify numerical data in a document.

3.2 Data Flagging

Data flagging identifies the existence of useful numerical data in documents and literature. Data flagging has been used in the INIS as a mechanism to create bibliographical data [7].

Generally, documents should be data flagged by performing two steps [8]: 1) add one or more of the narrower terms of the descriptor NUMERICAL DATA to the already chosen set of descriptors and 2) place the Literary Indicator tag "N," which emphasizes certain literary characteristics of a record, in the appropriate place within the bibliographic data elements.

We have considered such a method to identify numerical data and have attempted to implement a data flagging system in the INIS because data flagging is effective in searching valid numerical data. However, currently, using keywords in the INIS database is not possible. To implement data flagging to the Fukushima Nuclear Accident Archive, we must consider appropriate keyword methods and use tags that indicate numerical data.

3.3 Keywords for Data Flagging

Keywords for data flagging are very useful in the identification of numerical data. We have considered using keywords in the INIS database and the INSEPC database [9]. These databases each have a control thesaurus of terms related to nuclear and physics fields and numerical data indexing rules that include data flagging: categorized keywords (e.g., compiled, evaluated, experimental, and theoretical data.) in the INIS database, and physical quantity keywords (e.g., temperature, pressure, radiation exposure, and radioactivity) in the INSPEC database.

However, there is a lack of keywords closely related to the Fukushima Accident. If we only make use of both thesauruses' keyword, we must add additional terms related to the Fukushima Accident. To address this lack of terms, we used the content of *Plant Data of Fukushima Daiichi Nuclear Power Station at the time of the Tohoku-Chihou-Taiheiyoku-Oki Earthquake* [9] because this publication contains a guide to relevant data about the Fukushima Accident, e.g., plant parameters, data of abnormal events (including alarm records), and transient recorder data.

We have also checked the data in the Fukushima Nuclear Accident Archive because these data are closely related to the Fukushima Accident, such as water level, status of water leakage, radioactivity level, reactor vessel pressure, monitoring, and water injection into the reactor.

We considered the keywords listed in Fig. 5 for data flagging by combining the thesaurus **keywords** of the INIS and INSEPC databases and the content and terms extracted from the Fukushima Nuclear Accident Archive.

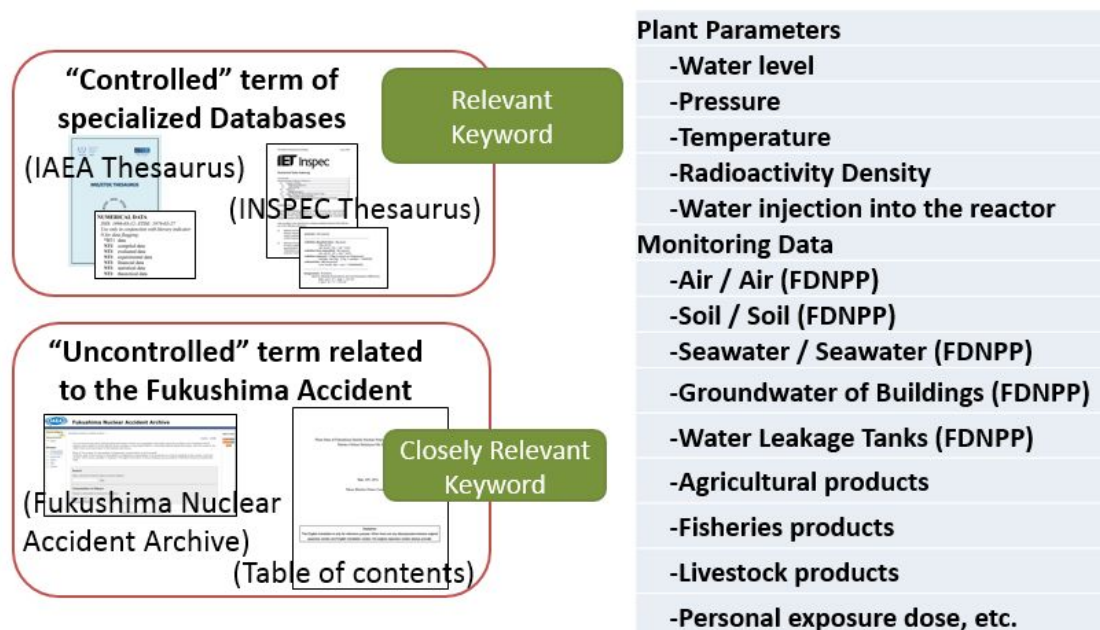


Fig. 5 Consideration of the keywords for data flagging on the basis of the two databases' thesaurus

3.4 Data Flagging in the Fukushima Nuclear Accident Archive

We use Simple Dublin Core as our metadata format and employ the bibliographical control rule for Internet information in the Fukushima Nuclear Accident Archive.

We use the following for data flagging [Fig. 6]:

- We assign "Numerical data" to the Type field, which indicates that the record contains numerical data.
- We assign keywords to the Subject.chapter field that indicate the type of numerical data.

Element	Value (ex)
Identifier.uri	http://warp.ndl.go.jp/info:ndljp/pid/3531775/www.nisa.meti.go.jp/english/files/en20110503-2-4.pdf
Title	Fukushima Dai-ichi Monitoring Data 2011-05-03 17:00
Format.extent	4p.
Format.medium	application/pdf
Type	Add Tag: "Numerical Data"
Date.issued	2011-05-03
Language	eng
Creator	METI
Subject.chapter	Add Keywords: "Monitoring Data-Air" "Monitoring Data-Air (FDNPP)" etc.

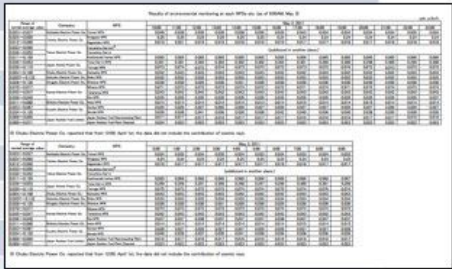


Fig. 6 Data flagging in the Fukushima Nuclear Accident Archive

Utilizing the tags and data flagging keywords, narrowing search results to those that contain numerical data is possible, which is beneficial for people who would like to access only numerical data.

3.5. Using Numerical Data Related to the Fukushima Accident

We face three issues in the development of effective metadata for numerical data related to the Fukushima Accident.

First, to create effective metadata, we must add multiple dates or times, the location of the observed or collected data, and distinguish between raw data and processed data. As a result, data can be obtained more accurately. However, despite using Simple Dublin Core to describe date information, multiple dates, such as date created, date modified, date collected, and date evaluated, cannot be distinguished because only one date field exists.

Second, to increase the convenience of the numerical data, conversion of various file formats (e.g., PDF, images, and graphs) to CSV or text format is required. However, this requires significant time and effort because such file formats represent non-structured data, the format of the data is frequently scattered, and the data is often not represented according to a unified format rule after initial conversion.

Third, we have no criteria to evaluate the validity and reliability of numerical data. In future, we plan to continue working on these challenging issues.

4. Conclusion

The amount of valuable numerical data related to the Fukushima Accident is increasing day by day, especially on the Internet. However, identifying whether such information contains numerical data even though we use search engines (e.g., Google) is difficult because numerical data can be provided in image files that are not indexed and websites do not always contain words that indicate numerical data.

We attempted to introduce a novel data flagging system to the Fukushima Nuclear Accident Archive to facilitate effective search for numerical data. Thus, users can significantly benefit by appropriate identification of the content of Internet information related to the Fukushima Accident.

We believe that our work will improve access to numerical data related to the Fukushima Accident and will hopefully contribute to improved circulation of grey literature from Japan to the world.

References

- [1] Ikeda, K., Ohshima, T., Gonda, M., Nagaya, S., Hayakawa, M., Kunii, K., Mineo, Y., Yonezawa, M., and Itabashi, K. (2014) 'Contribution to the improvement of dissemination of grey literature -JAEA Library's efforts for collecting, organizing and disseminating information on nuclear accidents- 'Proceedings of the Fifteenth International Conference of Grey Literature, pp.41-47.
- [2] Ikeda, K., Nakajima, H. and Yonezawa, M. (2012) 'Introduction of Useful Bibliographies and Resources related to the Fukushima Daiichi Nuclear Power Station Accident through the JAEA Library', ATOMOSU (ATOMO-sigma, Journal of the Atomic Energy Society of Japan), Vol.54, No.8, pp.549-553 (in Japanese).
- [3] Nakajima, H., Ikeda, K., Yonezawa, M., Itabashi, K., Kiriya, E. and Iwata, S. (2012) 'Study for development of the Fukushima Dai-ichi nuclear power station accident archive', Journal of Japan Society of Information and Knowledge, Vol.22, No.4, pp.344-353 (in Japanese), https://www.jstage.jst.go.jp/article/jsik/22/4/22_22_344/pdf, retrieved 30 June 2013.
- [4] Gonda, M., Ikeda, K., and Nagaya, S. (2014) 'Road to archiving of information relating to the Fukushima Nuclear Accident by the Japan Atomic Energy Agency Library(<Special feature>Earthquake disaster archives)', The journal of Information Science and Technology Association, Vol.64, No.9, pp.357-360 (in Japanese).
- [5] Fukushima Nuclear Accident Archive. <http://dspace.iaea.go.jp/dspace/?locale=en>
- [6] NDL Web Archiving Project (WARP).http://warp.da.ndl.go.jp/info/WARP_en.html
- [7] INIS Collection Search. <http://inis.iaea.org/search/>
- [8] INIS/ETDE Manual for Subject Analysis (2012) IAEA-INIS/ETDE-03, International Atomic Energy Agency.
<http://www.iaea.org/inis/products-services/publications/IAEA-INIS-ETDE-03/IAEA-INIS-ETDE-03-2012-08.pdf>
- [9] INSPEC. <http://thomsonreuters.com/inspec/>
- [10] Plant Data of Fukushima Daiichi Nuclear Power Station at the time of the Tohoku-Chihou-Taiheiyoku-Oki Earthquake, Tokyo Electric Power Company, May 16th, 2011.
http://warp.da.ndl.go.jp/info:ndljp/pid/6019786/www.tepco.co.jp/en/nu/fukushima-np/plant-data/f1_1_Setsumei.pdf

The Use of Web-scraping Software in Searching for Grey Literature

Neal R. Haddaway (Sweden)

Abstract

Searches for grey literature can require substantial resources to undertake but their inclusion is vital for research activities such as systematic reviews. Web scraping, the extraction of patterned data from web pages on the internet, has been developed in the private sector for business purposes, but it offers substantial benefits to those searching for grey literature. By building and sharing protocols that extract search results and other data from web pages, those looking for grey literature can drastically increase their transparency and resource efficiency. Various options exist in terms of web-scraping software and they are introduced herein.

The Challenge of Searching for Grey Literature

The editorial scrutiny and peer-review that form integral parts of commercial academic publishing are useful in assuring reliability and standardised reporting in published research. However, publication bias can cause an overestimation of effect sizes in syntheses of the (commercially) published literature (Gurevitch and Hedges 1999; Lortie et al. 2007). In a recent study by Kicinski et al. (2015), the largest analysis of publication bias in meta-analyses to-date, publication bias was detected across the Cochrane Library of systematic reviews, although there was evidence that more recent research suffered to a lesser degree, thanks to mitigation measures applied in medical research in recent decades.

Some applied subject areas, such as conservation biology, are particularly likely to be reported in sources other than academic journals, so called *practitioner-held data* (Haddaway and Bayliss in press); for example reports of the activities of non-governmental organisations. Such grey literature is vital for a range of research, policy and practical applications, particularly informing policy decision-making. Documents produced by governments, business, non-governmental organisation and academics can provide a range of useful information, but are often overlooked in traditional meta-analyses and literature reviews.

Systematic reviews were established in the medical sciences to collate and synthesise research on particular clinical interventions in a reliable, transparent and objective manner, and were a response to the susceptibility to bias common to traditional literature reviews (Allen and Richmond 2011). In the last decade systematic review methodology has been translated into a range of other subjects, including social science (Walker et al. 2013) and environmental management (CEE 2013). A key aspect of systematic review methodology is that searches are undertaken for grey literature to mitigate possible publication bias and to include practitioner-held data. These searches may fail to find any research that is ultimately included (e.g. Haddaway et al. 2014), but it is important for the reliability and transparency of the review to demonstrate that this is the case: other reviews have demonstrated significant proportions of grey literature in the synthesised evidence base (Bernes et al. 2015).

Systematic review searches for grey literature can be particularly challenging and time-consuming. No comprehensive database resources exist in the environmental sciences for grey literature, as in many other disciplines, and so searches must include web-based search engines, specialist databases such as repositories for theses, organisational web sites such as non-governmental organisations, governmental databases and university repositories. Typically between 30 (Pullin et al. 2013) and 70 (Haddaway et al. 2014) individual web sites are searched. Systematic reviews often complement these manual searches using web-based search engines; both general (e.g. Google) and academic (e.g. Google Scholar). Not only are searches of this number of resources time-consuming, but they are also typically undertaken in a very non-transparent manner: excluded articles are rarely recorded and searches are not readily updatable or repeatable. Furthermore, included resources must be listed individually by hand in any documentation of search activities, whilst search results from academic databases, such as Web of Science, can be downloaded as full citations.

Web Scraping Software: a potential solution

Data scraping is a term used to describe the extraction of data from an electronic file using a computer program. *Web scraping* describes the use of a program to extract data from HTML files on the internet. Typically this data is in the form of patterned data, particularly lists or tables. Programs that interact with web pages and extract data use sets of commands known as *application programming interfaces* (APIs). These APIs can be 'taught' to extract patterned data from single web pages or from all similar pages across an entire web site. Alternatively, automated interactions with websites can be built into APIs, such that links within a page can be 'clicked' and data extracted from subsequent pages. This is particularly useful for extracting data from multiple pages of search results. Furthermore, this interactivity allows users to automate the use of websites' search facilities, extracting data from multiple pages of search results and only requiring users to input search terms rather than having to navigate to and search each web site first.

One major current use for web scraping is for businesses to track pricing activities of their competitors: pricing can be established across an entire site in relatively short time scales and with minimal manual effort. Various other commercial drivers have caused a large number and variety of web scraping programs to have been developed in recent years (see Table 1). Some of these programs are free, whilst others are purely commercial and charge a one off or regular subscription fee.

These web scraping tools are equally as useful in the research realm. Specifically, they can provide valuable opportunities in the search for grey literature, by: i) making searches of multiple websites more resource-efficient; ii) drastically increasing transparency in search activities; and iii) allowing researches to share trained APIs for specific websites, further increasing resource-efficiency.

A further benefit of web scraping APIs relates to their use with traditional academic databases, such as Web of Science. Whilst citations, including abstracts, are readily extractable from most academic databases, many databases hold more useful information that is not readily exportable, for example corresponding author information. Web scraping tools can be used to extract this information from search results, allowing researchers to assemble contact lists that may prove particularly useful in requests for additional data, calls for submission of evidence, or invitations to take part in surveys, for example.

Table 1. List of Major web scraping tools

(adapted from <http://scraping.pro/software-for-web-scraping/>). Prices were accurate at the time of publication.

Platform	Description ^a	Cost ^b	URL
Import.io	"Instantly Turn Web Pages into Data. No Plugin, No Training, No Setup. Create custom APIs or crawl entire websites using our desktop app - no coding required!"	Free (charge for premium service)	www.import.io
DataToolBar	"The Data Toolbar is an intuitive web scraping tool that automates web data extraction process for your browser. Simply point to the data fields you want to collect and the tool does the rest for you."	Free to try (no export facility), \$24 for full version	www.datatoolbar.com
Visual Web Ripper	"Visual Web Ripper is a powerful visual tool used for automated web scraping, web harvesting and content extraction from the web. Our data extraction software can automatically walk through whole web sites and collect complete content structures such as product catalogs or search results."	\$299 (including 1 year of maintenance)	www.visualwebripper.com
Helium Scraper	"Extract data from any website. Choose what to extract with a few clicks. Create your own actions. Export extracted data to a variety [sic] of file formats. "	Basic version \$99 to Enterprise version \$699	www.heliumscraper.com
OutWit Hub	"OutWit Hub breaks down Web pages into their different constituents. Navigating from page to page automatically, it extracts information elements and organizes them into usable collections."	Lite version free, Pro version at \$89.90	www.outwit.com
Screen Scraper	"Screen Scraper automates copying text from a web page, clicking links, entering data into forms and submitting them, iterating through search results pages, downloading files (PDF, MS Word, images, etc.)."	Basic edition free, commercial versions from \$549 to \$2,799	www.screen-scraper.com
Web Content Extractor	"Web Context Extractor is a professional web data extraction software designed not only to perform the most of [sic] dull operations automatically, but also to greatly	\$89	www.newprosoft.com/web-content-extractor.htm

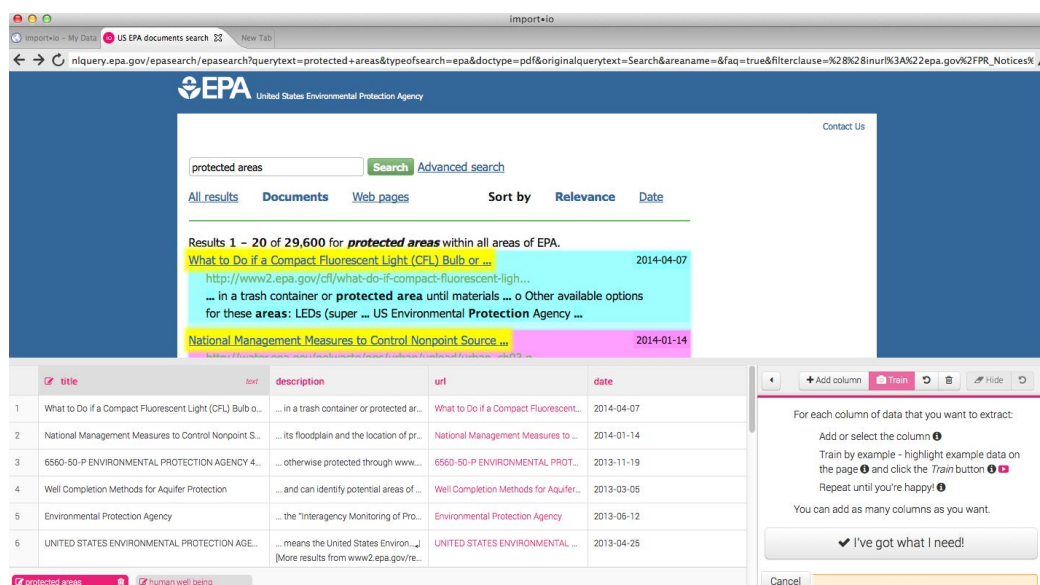
	increase productivity and effectiveness of the web data scraping process. Web Content Extractor is highly accurate and efficient for extracting data from websites.”		
Kimono	“Kimono lets you turn websites into APIs in seconds. You don’t need to write any code or install any software to extract data with Kimono. The easiest way to use Kimono is to add our bookmarklet to your browser’s bookmark bar. Then go to the website you want to get data from and click the bookmarklet. Select the data you want and Kimono does the rest.”	Free with additional features costing up to \$180	www.kimonolabs.com
FMiner	“FMiner is a software for web scraping, web data extraction, screen scraping, web harvesting, web crawling and web macro support for windows and Mac OS X. It is an easy to use web data extraction tool that combines best-in-class features with an intuitive visual project design tool, to make your next data mining project a breeze.”	Free 15 day trial, \$168-£248	www.fminer.com
Data Extractor by Mozenda	“The Mozenda data scraper tool is very basic; all you have to do is use the program to scrape up information you need off of [sic] websites without all the tiring work of searching websites one by one. Whether you are working for the government such as a police officer or a detective, in the medical field, or even a large business or entrepreneur, website scraping is fast, easy and affordable, plus it saves you or your employees a ton of stressful work and time; use Mozenda’s data scraper and let the program do all the hard work for you.”	Free trial (500 page credits), \$99 to \$199 per month	www.mozenda.com/data-extractor
WebHarvy Data Extractor Tool	“WebHarvy is a visual web scraper. There is absolutely no need to write any scripts or code to scrape data. You will be using WebHarvy's in-built browser to navigate web pages. You can select the data to be scraped with mouse clicks.”	\$99 - \$399	www.webharvy.com
Web Data Extractor	“Web Data Extractor [is] a powerful and easy-to-use application which helps you automatically extract specific information from web pages which is necessary in your day-to-day internet / email marketing or SEO activities. Extract targeted company contact data (email, phone, fax) from web for responsible b2b communication. Extract url, meta tag (title, desc, keyword) for website promotion, search directory creation, web research.”	\$89 - \$199	www.webextractor.com
Easy Web Extractor	“An easy-to-use tool for web scrape solutions (web data extracting, screen scraping) to scrape desired web content (text, url, image, html) from web pages just by few screen clicks. No programing required.”	\$69.99 (with in-app upgrades)	www.webextract.net
WebSundew	“WebSundew is a powerful web scraping tool that extracts data from the web pages with high productivity and speed. WebSundew enables users to automate the whole process of extracting and storing information from the web sites. You can capture large quantities of bad-structured data in minutes at any time in any place and save results in any format. Our customers use WebSundew to collect and analyze the wide range of data that exists on the Internet related to their industry.”	\$69 - \$2,495	www.websundew.com
Handy Web Extractor	“ Handy Web Extractor is a simple tool for everyday web content monitoring. It will periodically download the web page, extract the necessary content and display it in the window on your desktop. One may consider it as the data extraction software, taking its own niche [sic] in the scraping software and plugins.”	Free	www.scraping.pro/handy-web-extractor

^a Descriptions are taken from product web sites (31/01/2015)

^b Costs correct at time of publication (2015)

Figure 1 shows a screen shot of one web scraping program being used to establish an API for an automated search for grey literature from the website of the US Environmental Protection Agency. This particular web-scraping platform is in the form of a downloadable, desktop-based program; a web browser. The browser is then used to visit and train APIs by identifying rows and columns in the patterned data: in practice rows will typically be search records, whilst columns will be different aspects of the patterned data, such as titles, authors, sources, publication dates, descriptions, etc. Detailed methods for the use of web scrapers are available elsewhere (Haddaway et al. in press). In this way, citation-like information can be extracted for search results according to the level of detail provided by the website. In addition to extracting search results, as described above, static lists and individual, similarly patterned pages can also be extracted. Furthermore, active links can be maintained, allowing the user to examine linked information directly from the extracted database.

Figure 1. Screenshot of web scraping software being used to train an API for searching for grey literature on the Environmental Protection Agency website. Program used is Import.io.



Just as search results from organisational websites can be extracted as citations, as described above, search results from web-based search engines can be extracted and downloaded into databases of quasi-citations. Microsoft Academic Search

(<http://academic.research.microsoft.com>) results can be extracted in this way, and in fact a pre-trained API is available from Microsoft for extracting data from search results automatically (<http://academic.research.microsoft.com/about/Microsoft%20Academic%20Search%20API%20User%20Manual.pdf>).

Perhaps a more comprehensive alternative to Microsoft Academic Search is Google Scholar (<http://scholar.google.com>). Google Scholar, however, does not support the use of *bots* (automated attempts to access the Google Scholar server), and repeated querying of the server by a single IP address (approximately 180 queries or citation extractions in succession) can result in an IP address being blocked for an extended period (approximately 48-72 hours) (personal observation)¹. Whilst it is understandable that automated traffic could be a substantial problem for Google Scholar, automation of activities that would otherwise be laboriously undertaken by hand is arguably of great value to researchers with limited resources. Thus, a potential work-around involves the scraping of locally saved search results HTML pages after they had been downloaded individually or in bulk (this may still constitute an infringement of the Google Scholar conditions of use, however). A further cautionary note relates to demands on the servers that host the web sites being scraped. Scraping a significant volume of pages from one site or scraping multiple pages in a short period of time can put significant strain on smaller servers. However, the level of scraping necessary to extract 100s to 1,000s of search results is unlikely to have detrimental impacts on server functionality.

¹ Details of Google Scholar's acceptable use policy are available from the following web page: <https://scholar.google.co.uk/intl/en/scholar/about.html>.

Systematic reviewers must download hundreds or thousands of search results for later screening from a suite of different databases. At present, Google Scholar is only cursorily searched in most reviews (i.e. by examining the first 50 search results). The addition of Google Scholar as a resource for finding additional academic and grey literature has been demonstrated to be useful for systematic reviews (Haddaway et al. in press). Automating searches and transparency documenting the results would increase transparency and comprehensiveness of the reviews with a highly resource-efficient activity at little additional effort for reviewers. These implications apply equally to other situations where web-based searching is beneficial but potentially time-consuming.

Web scrapers are an attractive technological development in the field of grey literature. The availability of a wide range of free and low-cost web scraping software provides an opportunity for significant benefits to those with limited resources, particularly researchers working alone or small organisations. Future developments will make use of the software even easier; for example the one click, automatic training provided by Import.io (<https://magic.import.io>). Web scrapers can increase resource-efficiency and drastically improve transparency, and existing networks can benefit through readily sharable trained APIs. Furthermore, many programs can be easily used by those with minimal or no skill or prior knowledge of this form of information technology. Researchers could benefit substantially by investigating the applicability of web scraping to their own work.

Acknowledgments

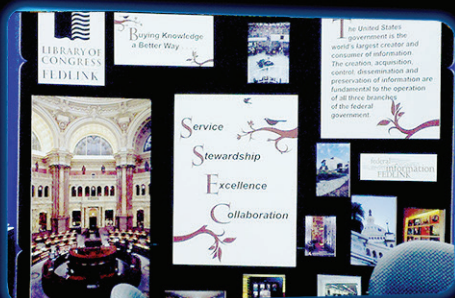
The author wishes to thank MISTRA EviEM for support during the preparation of this manuscript.

References

- Allen C, Richmond K (2011) The Cochrane Collaboration: International activity within Cochrane Review Groups in the first decade of the twenty-first century. *Journal of Evidence-Based Medicine* 4(1):2–7
- Bernes, C., Carpenter, S.R., Gårdmark, A., Larsson, P., Persson, L., Skov, C., Speed, J.D.M., Van Donk, E. (2015). What is the influence of a reduction of planktivorous and benthivorous fish on water quality in temperate eutrophic lakes? A systematic review. *Environmental Evidence*
- Collaboration for Environmental Evidence. 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. Environmental Evidence. Available from www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf (accessed January 2014).
- Gurevitch, J. and Hedges, L.V. 1999. Statistical issues in ecological meta-analyses. – *Ecol.* 80: 1142-1149.
- Haddaway, N.R. and Bayliss, H.R. 2015. Shades of grey: two forms of grey literature important for conservation reviews. In press
- Haddaway, N. R., Burden, A., Evans, C., Healey, J. R., Jones, D. L., Dalrymple, S. E., Pullin, A. S. (2014) Evaluating effects of land management on greenhouse gas fluxes and carbon balances in boreo-temperate lowland peatland systems. *Environmental Evidence*, 3:5.
- Haddaway, N.R., Collins, A.M., Coughlin, D., Kirk, S. (2015) The role of Google Scholar in academic searching and its applicability to grey literature searching. *PLOS ONE*, in press.
- Kicinski, M., Springate, D. A., Kontopantelis, E. (2015). Publication bias in meta - analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, 34: 2781-2793.
- Lortie, C.J. 2014. Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. – *Oikos* 123: 897-902.
- Pullin, A. S., Bangpan, M., Dalrymple, S. E., Dickson, K., Haddaway, N. R., Healey, J. R., Hauari, H., Hockley, N., Jones, J. P. G, Knight, T., Vigurs, C., Oliver, S. (2013) Human well-being impacts of terrestrial protected areas. *Environmental Evidence*, 2:19.
- Walker, D., G. Bergh, E. Page, and M. Duvendack. 2013. Adapting a Systematic Review for Social Research in International Development: A Case Study from the Child Protection Sector. London: ODI.

FEDLINK

an organization
of federal agencies
working together
to achieve optimum use of
resources and facilities
of federal libraries
and information centers
by promoting
common services,
coordinating and sharing
available resources, and
providing continuing
professional education.



Strategic Sourcing



Currently, more than 20 federal agencies – both military and civilian – including the Department of Defense, the Department of Justice, the Department of Health and Human Services, the Department of Education, the Department of the Interior, the Department of Agriculture, the Department of Commerce, the Department of Energy, the Department of Housing and Urban Development, the Department of Labor, the Department of State, the Department of Transportation, the Department of Veterans Affairs, the Department of the Treasury, the Office of Management and Budget, and the General Services Administration participate in the Federal Strategic Sourcing Initiative (FSSI). The FSSI was created in 2005 by the Department of the Treasury, the Office of Management and Budget, and the General Services Administration to identify products and services that can be purchased more efficiently through a centralized acquisition process. FSSI agencies also provide centralized acquisition functions for a variety of products to streamline efficiency and reduce costs to the federal government.



101 Independence Ave, SE ~ Washington, DC 20540-4935
FEDLINK Main Number (202) 707-4800
FEDLINK Hotline (202) 707-4900

Seventeenth International Conference on Grey Literature 'A New Wave of Textual and Non-Textual Grey Literature'



Royal Netherlands Academy of Arts and Sciences
Amsterdam, Netherlands on December 1-2, 2015

Call for Posters

Guidelines for Abstracts

Title of Poster:

Topic/Theme:

Author Name(s):

Telephone:

Organization(s):

Fax:

Address/P.O. Box:

Email:

Postal Code - City - Country:

URL:

Participants who seek to present a poster during GL17 are invited to submit an English abstract between 350-400 words. The abstract should describe the project and/or related information product or service. The abstract should likewise include a title, name(s) of the presenter(s) and their full address information. Abstracts are an important source of information available prior to the conference that is accessible to both delegates and conference participants.

Conference Related Topics

- ☐ New Types of Grey Literature Documents
- ☐ New Forms of Non-Textual Grey Literature
- ☐ The Convergence of Textual and Non-Textual Grey Literature
- ☐ The Influence of Social Media and Networks on Grey Literature
- ☐ Innovative ways in Leveraging Grey Literature
- ☐ Other Conference Topics

Due Date and Format for Submission

Timely registration is a guarantee of your placement on the conference program. The conference venue is only able to accommodate a limited number of physical posters (size A1). Abstracts in MS Word should be emailed to conference@textrelease.com on or before **October 15th 2015**. The author will receive verification upon receipt and this will be accompanied by further guidelines for submission of PowerPoint slides, Conference Registration, etc.

TextRelease

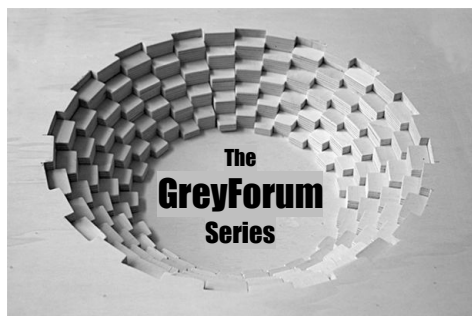
GL17 Program and Conference Bureau

JAVASTRAAT 194-HS, 1095 CP AMSTERDAM, THE NETHERLANDS
www.textrelease.com • conference@textrelease.com

Tel/Fax +31-20-331.2420

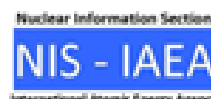
"The GreyForum is a series of onsite and online courses, seminars, and workshops where grey literature provides common ground for information professionals in the process of knowledge transfer"

Grey Literature and Digital Preservation: Standards in Practice



This pre-conference workshop is organized by GreyNet International together with speakers from Open Preservation Foundation (United Kingdom), NIS-IAEA (Austria), and CERN (Switzerland)

Grey Literature and Digital Preservation: Standards in Practice



GreyForum 4.1

Workshop

Grey Literature and Digital Preservation: Standards in Practice

Date

Monday, 30 November 2015

Time

10:30 – 15:00 hours

Location

Huis de Pinto
Literary and Cultural Center
Sint Antoniesbreestraat 69
1011 HB Amsterdam
The Netherlands

Registration

The Program includes coffee/tea service during the day and workshop materials. Early registration is recommended. Seating arrangements for twenty participants.

<http://www.greynet.org/greyforumseries/digitalpreservation.html>

Contact

GreyNet International
Grey Literature Network Service
Javastraat 194-HS, 1095 CP
Amsterdam, Netherlands

info@greynet.org
<http://www.greynet.org>
Phone/Fax +31 (0) 20 331 2420

About the Program and Panel:

Open Community Approaches to Digital Preservation

Ross King, Open Preservation Foundation, OPF, United Kingdom

Digitization and Preservation of Non-Conventional Literature for the International Nuclear Information System, INIS

Germain St-Pierre, Nuclear Information Section, IAEA, Austria

Research Data Analysis Preservation and Reuse: Case Study in High-Energy Physics

Tibor Simko, European Organization for Nuclear Research, CERN, Switzerland

Grey Literature and Digital Preservation is geared to service providers, publishers, librarians, archivists, and other information professionals in government, academics, business and industry involved in the curation and preservation of grey literature resources.

Program Outline

10:30 Arrival, Welcome	13:00 Lunch
10:45 OPF Presentation	14:00 Panel Discussion
11:30 NIS-IAEA Presentation	14:45 Take Away
12:15 CERN Presentation	15:00 Close

Note A special reduced fee is offered participants who are registered for the GL17 Conference in Amsterdam on December 1-2, 2015.

About the GreyForum Series www.greynet.org/greyforumseries.html

GreyNet

Börjesson, Lisa
137

Lisa Börjesson is a doctoral student at the Department of archival science, library & information science, and museum & heritage studies at Uppsala University, Sweden. She also holds a position as a visiting student researcher at UC Berkeley School of Information 2014-2015. Lisa has a background as a library development consultant working within regional government. She currently works together with archaeologists and information scholars in the research project Archaeological Information in the Digital Society (ARKDIS) funded by the Swedish Research Council, and is a member of the northwestern European Centre for Digital Heritage. Her ongoing dissertation research (2013-2017) explores three contextual factors shaping reporting in professional archaeology: information policy, report writers' frames of reference and their interpretations of the report writing as a professional task. Her research complements initiatives to improve technical access to archaeological grey literature by providing a more nuanced understanding for the report genre. Email: lisa.borjesson@abm.uu.se

Evans, Tim N.L.
145

Tim is responsible for accessioning, mounting and indexing the ADS (Archaeology Data Service) collections of electronic data, and the negotiation and validation of new deposits. Over the years he has taken a special interest in archaeological grey literature and has recently taken over the maintenance of OASIS. Prior to joining the ADS Tim worked in commercial archaeology for 6 years, primarily at Birmingham Archaeology (formerly BUFAU), working on a range of urban and rural sites across the West Midlands and beyond. He has also worked extensively overseas, including excavation and survey projects in Cyprus, Italy and Egypt. Email: tim.evans@york.ac.uk

Goggi, Sara
171

Sara Goggi is a technologist at CNR-ILC. She started working at ILC in 1996 working on the EC project LE-PAROLE for creating the Italian reference corpus; afterwards she began dealing with the management of several European projects and nowadays she is involved with organisational and managerial activities mainly concerning international relationships and dissemination as well as organization of events (e.g. LREC conference series). Currently one of her permanent activities is the editorial work for the international ISI Journal Language Resources and Evaluation, being its Assistant Editor. Since many years (from 2004) she also carries on research on terminology: this field of research starting intersecting with work on Grey Literature in 2011 at GL13.

Email: sara.goggi@ilc.cnr.it

Gonda, Mayuki
179

Mayuki Gonda works as librarian at the Central Library of JAEA (Japan Atomic Energy Agency). He joined JAERI (former JAEA) in 2005, and had been working for management and dissemination of JAEA research results information. Since 2009, he is in charge of selection, classification and indexing for INIS (International Nuclear Information System) Database. He is also a member of editorial committee of the Journal of Information Science and Technology Association (INFOSTA) since 2008. He holds a degree in information science (M.A.) from the Graduate School of Library, Information and Media Studies, University of Tsukuba.

Email: gonda.mayuki@jaea.go.jp

Haddaway, Neal R.
186

Neal Haddaway is a Project Manager at MISTRA EviEM based at the Royal Swedish Academy of Sciences in Stockholm. His research involves systematic reviews of environmental science research for use in science-policy.

Email: neal.haddaway@hotmail.com

Hayakawa, Misa
179

Misa Hayakawa works as librarian at the Central Library of Japan Atomic Energy Agency(JAEA). She Joined JAEA in 2010, and has been working for managing metadata of both papers and oral presentations published by JAEA researchers. In addition, she disseminates such information on the Internet via "JAEA Originated Papers Searching System"(JOPSS). She got the master's degree in Library and Information Science at the Graduate School of Library Information and Media Studies, University of Tsukuba in Japan.

Email: hayakawa.misa@jaea.go.jp

Ikeda, Kiyoshi
179

Kiyoshi Ikeda works as a librarian at the Central Library of Japan Atomic Energy Agency (JAEA). He also currently works as the chief of International Nuclear Information System (INIS) and Nuclear Information Management Section of JAEA. Since 2011, he has been in charge of selection and classification for INIS Database input data. He joined Japan Atomic Energy Research Institute (JAERI, former JAEA) in 2004, and has been working for acquisition of library materials (domestic/foreign books, journals, technical reports and proceedings). From 2007 to 2010, he had been working for editing work at "JAEA Reports" and "JAEA R&D Review" which are grey literatures. From 2008 to 2013, he has also been a member of editorial committee of the Japan Special Libraries Association (JSLA). He holds a degree in library and information science (M.A.) from the Graduate School of University of Library and Information Science (in Japan).

Email: ikeda.kiyoshi@jaea.go.jp

Pardelli, Gabriella
171

Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute since 1990, responsible for the Archives of the ILC Institute since 2005. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the Humanities. She has participated in many national and international projects including the recent projects:- BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council, (funded by CNR); - For digital edition of manuscripts of Ferdinand de Saussure (Research Programs of Relevant National Interest, PRIN - funded by the Ministry of Education, University and Research, MIUR). Email: gabriella.pardelli@ilc.cnr.it

Notes for Contributors

Non-Exclusive Rights Agreement

- I/We (the Author/s) hereby provide TextRelease (the Publisher) non-exclusive rights in print, digital, and electronic formats of the manuscript. In so doing,
- I/We allow TextRelease to act on my/our behalf to publish and distribute said work in whole or part provided all republications bear notice of its initial publication.
- I/We hereby state that this manuscript, including any tables, diagrams, or photographs does not infringe existing copyright agreements; and, thus indemnifies TextRelease against any such breach.
- I/We confer these rights without monetary compensation and with the understanding that TextRelease acts on behalf of the author/s.

Submission Requirements

Manuscripts should not exceed 15 double-spaced typed pages. The size of the page can be either A-4 or 8½x11 inches. Allow 4cm or 1½ inch from the top of each page. Provide the title, author(s) and affiliation(s) followed by your abstract, suggested keywords, and a brief biographical note.

A printout or PDF of the full text of your manuscript should be forwarded to the office of TextRelease. A corresponding MS Word file should either accompany the printed copy or be sent as an attachment by email. Both text and graphics are required in black and white.

REFERENCE GUIDELINES

General

- i. All manuscripts should contain references
- ii. Standardization should be maintained among the references provided
- iii. The more complete and accurate a reference, the more guarantee of an article's content and subsequent review.

Specific

- iv. Endnotes are preferred and should be numbered
- v. Hyperlinks need the accompanying name of resource and date; a simple URL is not acceptable
- vi. If the citation is to a corporate author, the acronym takes precedence
- vii. If the document type is known, it should be stated at the close of a citation.
- viii. If a citation is revised and refers to an edited and/or abridged work, the original source should also be mentioned.

Examples

Youngen, G.W. (1998), Citation patterns to traditional and electronic preprints in the published literature. - In: College & Research Libraries, 59 (5) Sep 1998, pp. 448-456. - ISSN 0010-0870

Crowe, J., G. Hodge, and D. Redmond (2010), Grey Literature Repositories: Tools for NGOs involved in public health activities in developing countries. – In: Grey Literature in Library and Information Studies, Chapter 13, pp. 199-214. – ISBN 978-3-598-11793-0

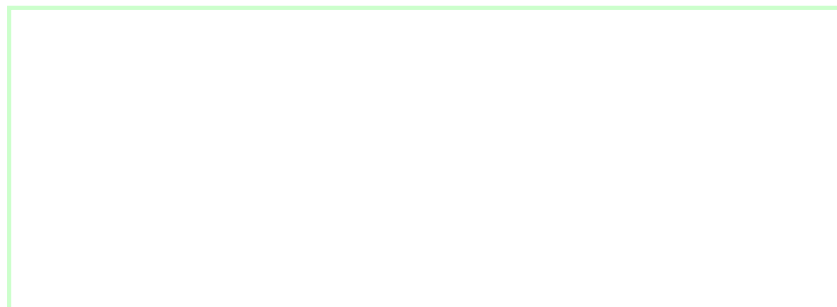
DCMI, Dublin Core Metadata Initiative Home Page http://purl.oclc.org/metadata/dublin_core/

Review Process

The Journal Editor first reviews each manuscript submitted. If the content is suited for publication and the submission requirements and guidelines complete, then the manuscript is sent to one or more Associate Editors for further review and comment. If the manuscript was previously published and there is no copyright infringement, then the Journal Editor could direct the manuscript straight away to the Technical Editor.

Journal Publication and Article Deposit

Once the journal article has completed the review process, it is scheduled for publication in The Grey Journal. If the Author indicated on the signed Rights Agreement that a preprint of the article be made available in GreyNet's Archive, then browsing and document delivery are immediately provided. Otherwise, this functionality is only available after the article's formal publication in the journal.



An International Journal on Grey Literature
'TOPICAL AND TECHNICAL ADVANCES IN GREY LITERATURE'
Autumn Issue, Volume 11, Number 3 2015

An Attempt to Nuance the Understanding of Professional Reports in Archaeology	137
<i>Lisa Börjesson (Sweden)</i>	
A Reassessment of Archaeological Grey Literature: semantics and paradoxes	145
<i>Tim N.L. Evans (United Kingdom)</i>	
Recent Italian regulations about the open-access availability of publicly funded research publications, and the documentation landscape in astrophysics	167
<i>Monica Marra (Italy)</i>	
Marine Planning and Service Platform; an advanced research engine for grey literature in marine science	171
<i>Sara Goggi, Monica Monachini, Francesca Frontini, Roberto Bartolini, Gabriella Pardelli, Maurizio De Mattei, Franco Bustaffa, and Giuseppe Manzella (Italy)</i>	
Enhancement of the functions of the Japan Atomic Energy Agency Library's Fukushima Nuclear Accident Archive using a novel data flagging system that improves the utilization of numerical data on the Internet	179
<i>Kiyoshi Ikeda, Mayuki Gonda, Shun Nagaya, Misa Hayakawa, Yukinobu Mineo, Katsuhiko Kuni, Minoru Yonezawa, and Keizo Itabashi (Japan)</i>	
The Use of Web-Scraping Software in Searching for Grey Literature	186
<i>Neal R. Haddaway (Sweden)</i>	

Colophon.....	134
Editor's Note.....	136
On the Newsfront	
GL17 Call for Posters : Seventeenth International Conference on Grey Literature.....	192
GreyForum 4.1 : Grey Literature and Digital Preservation.....	193
Advertisements	
EBSCO LISTA Full-Text.....	166
FEDLINK, Federal Library and Information Network - Library of Congress.....	191
Author information.....	194
Notes for Contributors.....	195