https://doi.org/10.26069/greynet-2026-000.548-tgj

ISSN 1574-1796

An International Journal on Grey Literature

Spring 2013 – TGJ Volume 9, Number 1

'Adapting New Technologies for Grey Literature'

www.textrelease.com GreyNet Grey Literature Network Service www.greynet.org

The Grey Journal
An International Journal on Grey Literature

COLOPHON

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 9, Number 1, Spring 2013. - *CVTISR, EBSCO, FEDLINK-Library of Congress, INIST-CNRS, ISTI-CNR, 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:
€125 individual, including postage & handling
€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/glpublications.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.

Errata
The cover of TGJ Volume 8, Number 3 should read Autumn 2012 instead of Summer 2012.

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.

Contents
'ADAPTING NEW TECHNOLOGIES FOR GREY LITERATURE'

https://doi.org/10.26069/greynet-2026-000.548-tgj

Editor's Note

The Fifteenth International Conference in the GL-Series provides the grey literature community an inclusive platform from which to assess developments in their field of information. Over the past two decades, since the very launch of this conference series, information science has been significantly impacted by social and technological developments. This gives sufficient cause for an audit in the field of grey literature – drawing upon accomplishments, assessing limitations, and projecting a sustained course of action.

A field assessment of grey literature extends well beyond library and information science, for it includes the assessment of grey literature produced and published in other sciences as well as government, business and industry. Information professionals and practitioners also become a part of this assessment, for it is they who carry out research in specific fields and make results available to their respective communities and wider public audiences.

The Grey Audit seeks to ascertain the validity and reliability of information and data produced in the grey circuit. It further seeks to measure the cost effectiveness of investing in grey literature both in material as well as human resources. The Grey Audit sets out to examine accepted standards applied in processing and distributing grey literature in an effort to identify guidelines for good practice that will be of benefit well into our 21st Century. Such examples of good practice will no doubt impact policy, which in turn will ensure future programs where grey literature is deployed.

In order to carry out The Grey Audit, information professionals with previous involvement in grey literature, as well as those new to the field are encouraged to respond to this year's Call for Papers.

Dominic Farace
journal@greynet.org

Grey Literature is a field in library and Information science that deals with the production, distribution, and access to multiple document types produced on all levels of government, academics, business, and industry in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body.

https://doi.org/10.26069/greynet-2026-000.548-tgj

https://doi.org/10.26069/greynet-2026-000.548-tgj

Creating and Assessing a Subject-based Blog for Current Awareness within a Cancer Care Environment *

Yongtao Lin and Marcus Vaska (Canada)

Abstract

The Health Information Network Calgary (HINC) is comprised of a group of libraries providing information services and resources to urban and rural sites in the Calgary Zone of Alberta Health Services. Establishing a current awareness service is a necessity in any discipline, especially in health care. Web 2.0 and social networks have transformed how health care professionals and researchers create knowledge, access information, collaborate, and disseminate research.

One of the earliest forms of social media, blogging has taken the world by storm (1) . Although there is a wealth of literature on the use of blogs in providing current awareness services for libraries, there is a pronounced gap on how blogs are assessed or evaluated, especially for information alert purposes (2) .

Clients within the HINC subscribe to e-mail alerts and RSS feeds, a trend particularly evident within the Cancer Care environment where a number of researchers have already implemented feed readers to remain aware of current literature. However, they often comment on challenges associated not only with maintaining alerts and managing RSS feeds, but also in selecting and creating alerts for unpublished materials. The need for a librarian-facilitated current awareness strategy became more and more apparent. The literature reviewed addressed the value of an alert, namely to indicate a gap in the participant's knowledge, rather than to deliver content the librarians may have perceived as useful (3). The authors saw the creation of a subject-based blog as an opportunity to disseminate current awareness "grey" information to this specific research community.

The Grey Horizon Blog was created in April 2012 using Blogger. The selection and re-aggregation of information involves ongoing assessment of user needs and continuous work on the Blog. A weekly global email-digest listing of the postings will be distributed two months after the launch.

Several metrics will be employed in October 2012 to evaluate the Blog. Blogger itself tracks the number of page-views over time. Google Analytics was set up as it tracks additional information on access and use of the Blog. As clients may be using feed readers to read Blog entries and may thus not visit the Blog at all, Feedburner has also been incorporated to track the number of times that the Blog RSS is accessed, as well as calculating the number of subscribers.

A post-survey will be conducted in six months to complement the web statistics data. The additional feedback and comments will help us determine whether the Blog has successfully created an easy platform for users to keep current with unpublished literature, the type of resources found most important, and whether the amount of time spent maintaining the Blog met expectations.

It is anticipated that this case study will portray how to successfully plan a subject-based blog to meet users' current awareness information needs in grey literature. Further efforts will focus on targeting the Blog to the topic areas in grey literature where users feel more information is needed. The findings from this assessment will direct us to potential marketing opportunities and changing technology that haven't been fully utilized in our Grey Horizon Blog.

Introduction: HINC and the Role of Social Media and Grey Literature in Healthcare

The Health Information Network Calgary (HINC) is a strategic partnership between the University of Calgary and Alberta Health Services (AHS). As a network of libraries covering different disciplines, the HINC is currently in the midst of progressing and extending coverage to a provincial-based service, collectively referred to as Knowledge Resource Services (KRS). The vision for the Network grew from a 2004 report, *Making Information Count*, advocating that "a transformed healthcare system needs a transformed information delivery system, if its practitioners, researchers, patients and their families are to have adequate and timely access to the information essential to their needs" (4). It proposed a new model for library services in the local health organizations, showcasing the successful integration of more dynamic information services supporting patient care, teaching, research, and learning, thereby strengthening the bond between healthcare practitioner and information provider (5). Of the six sites comprising the HINC, two cancer facilities serve nearly 500 cancer care staff.

As has been discussed time and time again, the power of grey literature lies in the fact that it is information that is rapidly produced, available when the user needs it, bypassing the often longwinded

* First published in the GL14 Conference Proceedings, February 2013.

peer-review process present in commercially published journals. In fact, this body of literature can perhaps best be equated with social media, in that it is "accessible to anyone at little or no cost…can have a very short time frame of production, and can be altered as needed" (1). With the advancement of technology, social media has become widely integrated in numerous aspects of information services. When considering the role of social media in the pursuit of grey literature, questions may arise as to whether or not healthcare professionals are prepared to accept this type of material as a means of keeping current and staying abreast of the latest information published in their disciplines. Social media engages readers with similar interests, fostering a sense of community development. In health care settings, social media has been widely adapted to "promote health, improve health care, and fight disease" (1). With *Grey Horizon* (http://grey-horizon.blogspot.com), social media has been adapted in our project as an effective platform for information and knowledge management. As such, it entails the provision of better communication, an alternative and expected social norm. As information professionals, we share our skills in retrieving and selecting appropriate grey literature material that will enhance our credibility in finding quality information.

One of the fundamental purposes of current awareness, exhibited in this paper by means of a blog, is to ensure the easy, convenient, and wherever possible, free access to information all in one place, offering support to the user every step of the way. As the discovery of grey literature material is heavily dependent upon information being made available online, engaging users to embrace social media components for their information-seeking pursuits is becoming crucial for health researchers. An interesting paradigm, considering that the blog, of which there are an estimated 70 million today, is often considered the earliest form of social media (1). According to a survey and research findings conducted in this field, 50% of physicians in Europe and 41% in the U.S. regularly engage and/or post in blogs (1). Further, 50% of medical organizations surveyed currently use blogs in their information pursuits, with another 80% planning to do so shortly (6). The cancer care librarians and creators of *Grey Horizon* thus echo Chu's notions that due to the interactive nature of the diary-like postings, "blogs are more dynamic and permit writers to engage in one to many conversations with their readers" (6).

Background: Current Awareness' Importance in Cancer Care Leads to Creation of the Blog

The notion of current awareness in the cancer care environment served by the *Grey Horizon* Blog is not new. Current Awareness services, "designed to keep users informed about recent developments" (7), have existed in Cancer Care since 2008. Librarians involved in this study have responded to user demand by spending countless hours over the years meeting with researchers and health care professionals on an individual basis to create current awareness guides or digests, providing instruction on how to set up a table of contents e-mail alert and/or an RSS feed for a particular journal or search strategy, catering to their clients' needs. While cancer researchers were keen, motivated, and engaged with this initiative at the beginning, increasing workloads, time constraints, and other pressing commitments soon put current awareness on the backburner. Many researchers echoed the sentiments of Attfield, Blandford, and Makri, as "participants frequently found attending to current awareness information overwhelming" (7). By sorting, selecting, and placing relevant information in a central location (i.e. blog) that all researchers could access on their own time (in lieu of filtering out a deluge of e-mails), the authors of this paper believe that issues of information overload and time constraints can be better managed.

For a librarian not to recognize his/her role as an information mediator and accept the direction that social media and the selective dissemination of information is taking, is to demonstrate that one's role in this profession is done. While the introduction of a blog in the specific cancer care environment discussed in this paper is a new initiative, it is not entirely unique, but rather merely goes with the flow of the connected health care professional. As a subject-based blog, *Grey Horizon* delivers timely and accurate information exclusively from grey literature resources, to the community in need.

Method: Designing and Creating the Grey Horizon Blog

Deciding on a subject-based blog

Findings from the literature support a strong association between current awareness and social media in healthcare. The HINC has and continues to integrate social media in an effort to reach out to the connected users served by this organization. Instant messaging chat reference, a *Twitte*r account, and a *YouTube* channel containing brief self-paced tutorials, all demonstrate the need of library services to recognize the importance of this communication medium. The nature of the cancer care environment within which the authors of this paper are employed is fast-paced with ongoing clinical trials, evidence-based guideline development, , grant proposal applications, physician meetings, and conference presentation opportunities. In fact, Health Information Network staff can be seen as information

brokers, acting "as a layer of 'intelligent filters' sensitive to complex, local information needs; their decisions aim to optimize conflicting constraints of recall, precision, and information quantity" (7). Cancer Care librarians integrate current awareness in their provision of information, by offering in-person training, facilitating virtual interest groups, and creating subject-specific e-mail alert digests, emphasizing that librarianship most go forward and acknowledge the joint role of current awareness and grey literature, particularly in a 21st century technologically-enriched society.

Planning for the Blog

Despite the wealth of information available focussing on the need and importance of blogs, far less literature has been written on evaluating and assessing a blog (2). The authors of this paper have thus applied Blair and Level's guide of subject blogging etiquette while planning *Grey Horizon*. Before embarking on the actual creation of the Blog, the authors first established and planned the process and methodology dictating how *Grey* Horizon would come to fruition. Blogger (http://www.Blogger.com) was chosen as the software, due to the program's ease of posting information, tracking followers, and generating statistics. Further, the program is free anduser-friendly. Once the program of choice was decided on, appropriate material from which the Blog's postings would be generated was selected, based on the authors' past experience and knowledge of grey literature cancer resources. This included a wide array of grey literature cancer resources, such as Canadian Partnership Against Cancer, Cancer Trials Canada, New York Academy of Medicine, Canadian Cancer Society, Canadian Health News, NICE guidelines, American Society of Clinical Oncology, American Association for Cancer Research, the European Society of Medical Oncology, and several others.

Organization of content was considered an essential planning step, as the volume of posts in a Blog can become unwieldy if not appropriately managed. To manage the scope and breadth of content, a controlled list, consisting of a series of tags corresponding to grey literature types, was established at the outset. Each new item posted on the Blog would thus be labelled accordingly. The most common tags, many of which were formed according to the nature of the postings, were Cancer Research, News, Conferences, Drug Updates, Case Studies, Reports, Clinical Trials, and Guidelines. In addition to the tagging mechanism utilized, the Blog's interface also contained a Favourite Links section, as well as an area where the reader could find out additional information about grey literature, as well as instructions on how to set up an e-mail alert and RSS feed.

To keep track of all of the information and make it easier for the project team to oversee all information sources for *Grey* Horizon, a shared email account was created via Gmail for storing and monitoring all RSS feeds and e-mail alerts. Meanwhile, a Blog log was initiated to keep track of the progress of postings, complete with time spent and any comments the Blog creators had with regards to issues encountered with the postings (i.e. website down, inappropriate posts, etc.) Inclusion and exclusion criteria were established before the Blog went live, to ensure that postings covered numerous aspects of cancer care, including research findings, clinical trials, and conference proceedings, as well as to maintain consistency among blog postings. Criteria common across all document types, and adversely affecting the decision of whether or not to post a particular news item included current, unbiased information from selected sources, clinical trials not endorsed by pharmaceutical companies, recognizing the various formats of grey literature and posting accordingly, as well as adhering to the main subject areas/types of sources previously identified from the users. To attract and sustain reader attention, new postings must appear on a daily basis.

Creation of the Blog

Grey Horizon was created in April 2012, and officially launched on April 30, 2012. Prior to the official unveiling, a soft launch was held for two weeks to allow for pretesting with a few key users, measure workload involved, and finalize the interface design.

Bi-weekly digests

Three months after launching the Blog, the librarians decided to re-aggregate selected postings into a bi-weekly digest format. The digest was disseminated to all staff at both cancer sites via an e-mail listserv. This promotion effort, combined with a link to the Blog being placed on the HINC website, established multiple avenues by which the Blog could be accessed and postings read. Along with the posting's relevancy to the cancer community at large, presenting information via a digest format provides a brief and condensed way for users to access the same information. Due to the nature of grey literature and the importance of disseminating current information in cancer care to Blog readers, *Grey Horizon* is reviewed daily, not only in posting up the latest cancer care news, but also in allowing our readers determine the direction the Blog will take. While the primary evaluation period was not scheduled until six months after the Blog launch date, librarians review postings and listen to reader comments on a continuous basis, thus following "what users would change or like to see added" (2).

Monitoring and maintaining the Blog

Due to the elusive nature of grey literature, websites that served as the focal point one day can be moved or disappear. Thus, as Attfield, Blandford, and Makri comment, the importance of continuous monitoring, particularly with a form of social media such as the blog, cannot be overlooked: "monitoring is an ubiquitous activity and can take many forms, frequently combining both formal routes with less formal routes, such as the use of social networks" (7). As creators of the Blog, the two cancer care librarians applied Ellis' model of information seeking (8-9); namely, appropriate cancer resources were identified and located (whether by conducting a search or implementing an e-mail alert or RSS feed), accessed, selected, and processed (i.e. posted on the Blog).

Methods: Analyzing and Evaluating the Blog

As there is no one standard metric for the evaluation of a blog (2), the authors decided to employ several metrics addressed in the literature to evaluate the usefulness and effectiveness of *Grey Horizon* . This included looking at participation through comments (10), webtrackers and other RSS feed reader tracking websites, in addition to monitoring blog traffic. Both qualitative and quantitative data from the post-survey and comments were gathered to assess how the blog meets successful criteria.

Blogger

Blogger, as an online platform for the creation of the blog, tracks the number of page-views over time. A total of 6806 page views on 463 postings were tracked from the Blogger from May to October 2012. As indicated from Figure 1, traffic has significantly increased during the six-month project phase. There were nearly three times (1835 page views) as many visitors in October as the first month in May (691 page views)

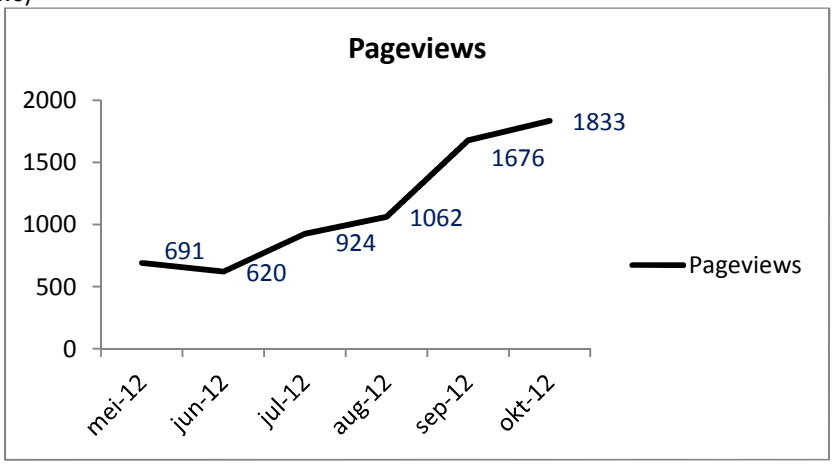


Figure 1: Number of page views by month from Blogger

The popularity of individual postings was also determined by the number of page views tracked, thus confirming the value of core types of grey literature for cancer researchers. Based on the total number of 463 postings as of October 22, 2012, the following was noted:

- 84 page views for news about a specific research **conference**
- 56 page views for a cancer **drug update** from Health Canada
- 53 page views for a **report** on cancer health services
- 51 page views for a clinical **guideline** from the National Institute for Health and Clinical Excellence (NICE)
- 46 page views for **statistics** on cancer incidence, survival and risk factors

Further, a definition of grey literature (66 page views), along with instructions on how to obtain full text papers from the references mentioned in the postings (23 page views) were among the most accessed pages within the More Information section of the Blog. Undoubtedly, users seek expert guidance in successfully retrieving, filtering, and understanding information.

Google Analytics

As a popular web tracking instrument, *Google Analytics* captures how visitors interact with a website. Therefore, this tool was set up to track additional information on access and use of the *Grey Horizon* Blog, including number of visits, number of hits, types of access, visitors from referring websites, and detailed user behaviour when visiting the Blog. The breakdown of new visitors and return visits was particularly helpful in evaluating the usefulness of the content and effectiveness of blog promoting strategies (Figure 2).

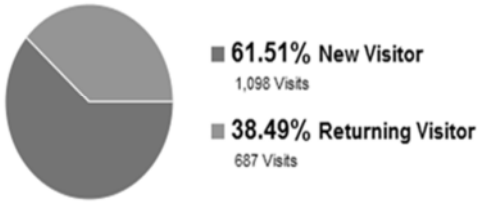


Figure 2: Percentage of new visitors vs. returning visitors of the Blog

According to *Google Analytics*, a total of 1785 people have visited *Grey Horizon*, as identified by unique IP address. An average visit duration of 4 minutes 14 seconds indicates that time is being taken to read and digest the information presented. In addition, it is helpful to examine how visitors navigate various pages, along with the different interactions undertaken.

Feedburner

Feedburner is a tracking service for RSS subscribers. As clients may be using feed readers to peruse blog entries and may thus not visit the Blog at all, by replacing the automatically generated RSS feed with one from *Feedburner*, *Feedburner* tracks the number of times that the Blog RSS is accessed, as well as calculating the number of subscribers. A total of 17 people subscribed to the *Grey Horizon* RSS feed over six months. On average, 12 out of 17 subscribers remained active by taking the actions of clicking the links to the postings, or viewing them in their feed readers. Responses from the post-survey questions on whether visitors subscribed to the Blog feed and their experience regarding use of an RSS feed reader confirmed users' education needs on using RSS feeds for current awareness purposes, a future instruction effort for librarians via more integrated alert services.

Reader feedback/comments

Having the ability and opportunity to comment based on postings is one of the most important features of a blog. The use of comments often foreshadow a blog's success. Since the "Comment" feature was turned on in September 2012, *Grey Horizon* has received seven comments from readers. Some echoed their own experiences with cancer, while others expressed their opinions on some of the research posted. Reader feedback and comments on our Blog were also observed from the connection activities with the Health Information Network Twitter account. The HINC Website links the *Grey Horizon* Blog in the current library news column and the tweets on individual postings were therefore tracked accordingly. As no single evaluation metric is able to present the full story, it was extremely helpful with ongoing service planning to have seen similar comments and feedback from different evaluation channels.

Post survey

A post-survey was distributed to the Cancer Care listserv in September 2012 to complement the web statistics data, with the purpose of helping the authors determine whether the Blog has successfully created an easy platform for users to keep current with unpublished literature, the type of resources found most important, and whether the amount of time staff spent maintaining the Blog met users' expectations. Fifty-one people completed the survey, comprised of questions in four sections:

- Users' familiarity with grey literature and current awareness practices
- Experience with the Blog
- Experience with bi-weekly digests
- Additional comments and suggestions on future current awareness services

43.5% of respondents mentioned unfamiliarity with the concept of grey literature prior to accessing the Blog. This finding was consistent with previous studies conducted by the authors indicating that most users successfully incorporated grey literature in their research despite being unaware of this term. (11) Google still appeared to be the most predominant source of information for grey literature searching for 21 out of 51 respondents. Despite an overwhelming number of people (67.3%) finding it challenging to keep abreast of current information in their fields, cancer researchers and healthcare professionals are still in the traditional mode of receiving and sharing new information. Attending workshops and conferences, setting up journal alerts via email, and communicating with colleagues are among the most frequent practices.

The post-survey has also reflected the success of the *Grey Horizon* Blog. 35 out of 51 respondents had accessed the Blog, a frequency ranging from a few times a month to every day. Among these, nearly all respondents (94.1%) found the postings useful. For those who hadn't accessed the Blog before, having

a busy clinical/research schedule was the most noted barrier. The usefulness of the topics was rated by the readers, as shown in Figure 3. Cancer Research, Clinical Trials, Reports and Drug Updates were the top four categories, the findings being consistent with the number of page views tracked by Blogger.

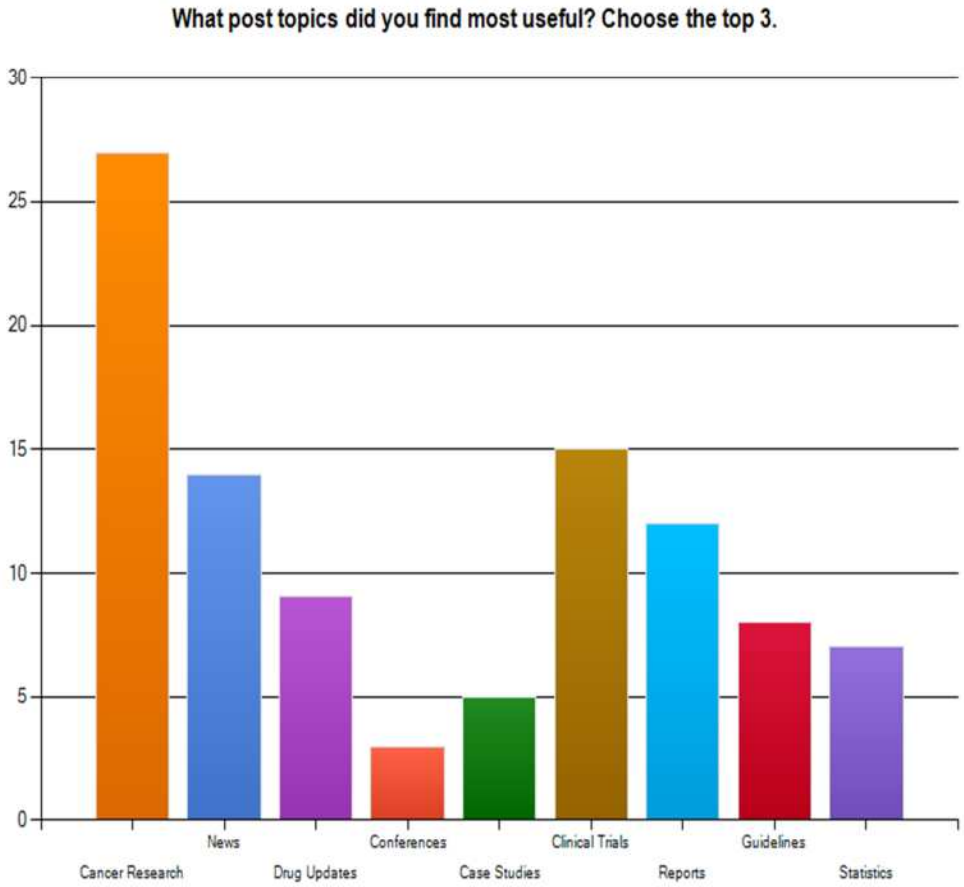


Figure 3. Usefulness of posting topics perceived from the Post-Survey

Additional features of the Blog, including providing links to other websites, pointing to the published studies mentioned in the post, and creating pages to inform about grey literature and how to access full-text articles were most favoured by the readers.

Bi-weekly digests appeared to be a very successful pursuit in re-aggregating the information in a user-friendly format, promoting the site to new staff as well as serving as reminders for existing readers to revisit the Blog. Ongoing assessment of the user needs and thorough planning of this current awareness project have become common themes the authors discovered when analyzing and evaluating the success of the Blog. When readers were asked to articulate their views on what may have contributed to the usefulness of the Blog and the postings, key features of a blog, including ease of interface navigation and information gathered in one place, as well as selection of various relevant sources and provision of current, pertinent information in one place were stated as the main criteria.

Discussion and Future Directions

Enhanced research interaction and collaboration

Grey literature's placement and availability on the World Wide Web, particularly the new Web 2.0 generation is essential to its success. These new services "encourage group interaction in a space where individuals can participate, socialize, and set social norms" (12). Such was the ideology behind the creation of *Grey Horizon*: to allow readers to not only learn about the latest cancer care news, but also to actively participate and interact with their colleagues, as well as with the cancer care librarians. The Blog has achieved a chain-reaction effect, demonstrating the power of collaborative learning and information sharing.

Librarians' integrated roles in current awareness services

Librarians in this project act as the conduits of information, browsing through RSS feeds and e-mail alerts to actively pull relevant information and responsively push this out to the users, the readers of the Blog (7). The authors are thus the gatekeepers, controlling the flow of information to craft Cancer Care

South into a well-informed and efficient organization (7). Although the creators of the *Grey Horizon* Blog identified resources that, in their opinion, were best suited to researchers within a cancer care research facility, it is important to note that the two librarians do not consider themselves as content experts in this field. This then emphasizes the collaboration needed between librarian and researcher, where the librarian seeks a better understanding of the information being requested is a form of contextual inquiry, playing a key role in the data that is gathered and eventually placed in the Blog (7).

Blog as a successful platform in promoting current awareness services

It is evident from this project that many researchers and healthcare professionals appreciated being "pushed" relevant information in their field, as they often did not have the time to go searching for this information themselves. Tailored information to different fields of interest in cancer was suggested as a future direction for a more dynamic and laddered current awareness service. The Blog has also become an invaluable tool that can aid in planning integrated information services, and promoting library instruction courses, special events, and new resources.

Becoming a source of grey literature

Although *Grey Horizon* has only been in existence for a mere 6 months, it is gaining ground and is serving as an important source of grey literature material in the cancer care environment in Alberta. This relative success within a short time frame demonstrates that subject-based blogs work well as a current awareness service, highlighting the importance of disseminating useful content on relevant topics to the intended audience. Readership is growing, requests from readers to link to external blogs is raising even further awareness.

Current awareness project planning

Grey Horizon reinforced the need for thorough project planning, achieving balance between staff time, workload, and user expectations. Additional features of the Blog are being explored, along with new methods of marketing and promotion. The authors are presently investigating the possibilities of linking and sharing with other blogs in the same subject field, as well as creating a mobile site for the Blog. . Despite making a few adjustments over the course of this pilot study, the creators of the Blog often heed the words of Chu et al. words that can be applied to the importance of intertwining current awareness with grey literature: "what attracts users is not technology but how you make use of the technology so that they can fully utilize the tools to accomplish things they want" (6).

References

(1) Graham DL. Social media and oncology: Opportunity with risk 2011:421-424.

(2) Blair J, Level AV. Creating and evaluating a subject-based blog: planning, implementation, and assessment. Reference Services Review 2008;36(2):156-166.

(3) Attfield S, Blandford A. Conceptual misfits in e-mail-based current-awareness interaction. Journal of Documentation 2011;67(1):33-55.

(4) Calgary Health Region, University of Calgary. Making Information Count: AN integrated knowledge service for healthcare practitioners, staff, patients and their families 2004.

(5) Reaume R, Aitken E, Powelson S. Library advocacy at the bedside and the boardroom: Calgary Health Information Network 2010; Available at: http://www.chla-absc.ca/2010/graphics/chla2010-poster17.pdf. Accessed October 29, 2012.

(6) Chu SKW, Woo M, King RB, Choi S, Cheng M, Koo P. Examining the application of Web 2.0 in medical-related organisations. Health Information & Libraries Journal 2011;29(1):47-60.

(7) Attfield S, Blandford A, Makri S. Social and interactional practices for disseminating current awareness information in an organisational setting. Information Processing and Management 2010;46(6):632-645.

(8) Ellis D. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. The Library Quarterly 1993;63(4):469-486.

(9) Meho LI, Tibbo HR. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. Journal of the American Society for Information Science and Technology 2003;54(6):570-587.

(10) Jackson A, Yates J, Orlikowski W. Corporate blogging: Building community through persistent digital talk. Proceedings of the 40th Hawaii International Conference on System Science; Available at http://doi.ieeecomputersociety.org/10.1109/HICSS.2007.155. Accessed November 12, 2012.

(11) Lin Y, Vaska M. Raising awareness of grey literature in an academic community using the cognitive behavioral theory. Eleventh International Conference on Grey Literature: The Grey Mosaic, Piecing it all Together; Available at http://www.opengrey.eu/item/display/10068/698105. Accessed November 14, 2012.

(12) Cho A. AN introduction to mashups for health librarians. Journal of the Canadian Health Libraries Association 2007;28:19-22.

https://doi.org/10.26069/greynet-2026-000.548-tgj

Tools and Resources Supporting Cultural Tourism*

Eva Sassolini, Alessandra Cinini, Stefano Sbrulli, and Eugenio Picchi (Italy)

Introduction

The diffusion of internet and the information technologies are creating continuous information flows. There is a widespread awareness of the added value and of the role that the Web has in dissemination, exploitation and promotion of the cultural tourism, especially in a country like Italy, where the cultural heritage is very important. Moreover, an open philosophy causes problems of authoritativeness in the production of contents because it is characterized by a strong interaction among users thus creating a distance between knowledge and communication. The spread of internet has brought the significant changes of communication paradigm. Nowadays the competition decreases among contents, even among from sources published in potential competition with them. In network logic, all nodes are interdependent and represent a single large hypertext. The proliferation of paths boosts a free circulation of ideas and can bring out most interesting contents[1].

The Projects

Methodologies, strategies, resources and tools created in our research group allowed us to take part in some national initiatives:

- "On-line dissemination of the historical artistic and landscaped, regional heritage" *(WeBasCH)*: The project was born within the framework of collaboration between the ILC-CNR Pisa and the APT Basilicata (i.e. Agenzia di Promozione Territoriale della regione Basilicata) to experiment and implement strategies for promotion and dissemination of regional heritage.
- *"SmartCity*: new solutions for content engineering and ambient intelligence as support of cultural tourism" is a Tuscany region project ("POR Creo" session), funded by the European Community (FESR). The specific project purpose is to allow important steps forward in terms of productivity, versatility and adaptability of digital content, both textual and multimedia, through knowledge engineering techniques. Particularly we have developed and tested tools for a better preservation and enhancement of cultural heritage, identifying methodologies and solutions to meet new demand of cultural spaces in particular for tourist purposes. Partners of this project were ILC-CNR and a consortium of companies: Space, Rigel Engineering and Meta.

The specific goal of both projects is experimenting and implementing strategies for promotion and dissemination of regional heritage through an effective communication style. By exploiting potentialities of the Web 2.0, we offered a set of chosen documents and cultural routes that increase knowledge of historic and landscaped resources. The aim is to join maintenance and preservation activities and enhancement and promotion of the cultural heritage initiatives with an increasing opening to the general user[2].

Text material for cultural tourism

The cultural tourism in Italy aims at identifying town of artistic interest or cultural tours in wider areas linked by historical events or traditions as the "Via Francigena", " Parco Archeologico Storico Naturale delle Chiese Rupestri ", "the wine roads", etc.. In this context, information available is in continuous evolution and so various that the database should be continually updated. The updating process of textual resources is an open question. The manually generation of resources is expensive and it requires the review of human experts. The manual approach is prone to errors of omission. Moreover, an approach based on an automatic updating of resources is not efficient because these are constantly evolving, this reason advises against an exclusive use of automatic applications. Especially when it comes to tourism, for which the Web offers constantly new ideas: a new park, an archaeological site discovered, etc..

There are many European initiatives within Cultural Heritage that aim to develop knowledge and enhancement of digital cultural heritage that have been undertaken. Among these, "Minerva" and "Michael" have been coordinated by the Italian Ministry for Cultural Heritage. Minerva has developed a platform of guidelines and recommendations, which are shared by European member states, for the digitization of cultural heritage and its network access. Similar information can be considered the reliable source of knowledge, especially because they were created with specific objectives:

* First published in the GL14 Conference Proceedings, February 2013.

- ✓ accessibility and visibility improvement of European digital cultural resources;
- ✓ support development of the European Digital Library for a better access to the cultural resources;
- ✓ contribution to increasing the interoperability among existing networks;
- ✓ promoting the use of digital cultural resources by business and citizens.

However the creation of this kind of resources is also referred to written text material such as brochures and web advertisement, that often is difficult to find, and retrieve. Such data regard an important source of information but since it tends to be original, recent and ephemeral can be considered a typical material of Grey Literature.

Specific strategies for texts acquisition

We have dealt the creation of the text corpus with different strategies in the two projects. Different needs of each project have guided the choices of the best acquisition strategy and all text material has collected in two steps:

1. We built a starting text corpus (hereafter C0) which was exploited to generate new linguistic resources and enriching the ones available from TextPower[i][3] project.
2. On the basis of these resources, we enlarged C0 to create the reference corpus.

CO in *SmartCity*

The first acquisition strategy in *SmartCity* project concerned the creation of a domain corpus "Empoli e dintorni", composed of historical, artistic and tourist documents related to the specific geographical area. Among activities of the project it has been possible collaborate with the library of Empoli and Cerreto Guidi for the retrieval of textual material.

On a first group of documents provided by the library of Empoli, we attempted to select material of interest in order to deal all the most important themes of Empoli and surroundings. On the basis of these texts we generated the Training Corpus whose size is almost 110,00 occurrences of words.

CO in *WeBasCH*

A different approach has been followed about *WeBasCH* project where all text material has retrieved on-line. The APT Basilicata provided us with a list of institutional websites. The items of the list more related to historical, artistic and landscaped regional heritage were selected and browsed by using automatic spiders and parsers, for the creation of a text corpus. Much of documentation available in internet can be assimilated to the "grey literature", since it has been produced by the authors and institutions outside publishing, particularly the websites of regional entities.

Text corpus to linguistic resources

C0 was exploited to generate new linguistic resources and integrating the ones already available.

All textual material obtained has indexed and, after a tagging phase with the PiTagger[ii] tool, we could identify all lemmas and relative POS in each document[4].

PiTagger associates each word to the related lemma by using the morphological component of the Italian language PiMorfo[iii]. Then it solves the ambiguities by following a statistical approach and with the help of a training corpus.

Later on, all *multiword* expression (MWE) were extracted from C0 by exploiting pattern matching techniques. Typically for the Italian syntactic construction, the most productive linguistic patterns are N-preposition-N and ADj-N/N-ADj. Statistical algorithms analyze the distributions frequency of each pattern identified. On the basis of results we extracted a set of semantically relevant terms and concepts for the cultural heritage domain. The analysis of the collected texts by means of linguistic tools (morphological engine and tagger) is fundamental for productive application of the statistical functions of extraction[5].

We exploited enriched text material, that composes C0, to build weighed domain lexicons besides. Starting from a small set of relevant pivot terms a lexicon is obtained by means of *mutual information* criteria. Statistical algorithms analyze and weigh the frequency of the co-occurrency of each word with the pivot terms. The domain lexicons can be used to evaluate the relevance of a document for that domain, in this case it is most important to establish a minimal threshold.

[i] Project made for building of terminological resources and enrichment and annotation of textual material

[ii] PiTagger is an important component for text lemmatization and tagging and constitutes a software module of PiSystem: integrated system for processing of textual and lexical materials.

[iii] PiMorfo: system for morphological analysis of the Italian language.

Reference (text) Corpus

Since textual documentation collected didn't cover exhaustively any events and cultural resources, we tried to increase the amount of textual material available, in order to find a maximum coverage of information and knowledge. In both projects we used a new search strategy of text material retrieval, namely, we used extracted knowledge and specialized crawlers that work on a bulk of text material available on-line. In a next phase we have developed other tools that, by using specific semantic filters, make a ranking the documents and evaluate their relevance with the specific domain. All extracted documents were joined in C0 corpus to build the reference (text) corpus.

The (text) reference Corpus created in SMARTCITY project consists of 2219 text units:

- 1634 units come from the material provided by the library of Empoli and Cerreto Guidi, about 1.000.000 (997299) words;
- 585 units are related to texts retrieved on the Web, little more than 850.00 words (857355).

At the end of the project the specific Corpus has been reorganized in 650 documents in XML format.
In the *WeBasCH* project, the (text) reference Corpus instead is constituted of almost 2 million and half words. As it described, the creation required two phases, in which we retrieved documents from the Web and built the reference (text) corpus.

DBT-Faccette

The intelligent browsing system of text, named "DBT Faccette", is a customization of the categorization system used in librarianship. Its uniqueness lies in possibility of exploiting the semantic relevant elements (or "micro semantics") identified in the text for suggesting a further search. This feature makes most important the availability of an annotated corpus.
The text corpus is originally a set of text files. Inside these texts specific tools have identified annotations of various kinds: information related to words, phrases and piece of text. However it is necessary to organize the annotated texts in a coherent way and orderly, for a better storage, management and expansibility over time.
The corpus becomes a collection of digital sources containing accurate semantic annotations and necessary information to managing of the textual sources.
A useful way to imagine a good browsing system is to assume the final usage scenarios. Many developers often tend to adopt "reference systems" implicit, made of the individual experiences. Since the "reference system" of a user is not necessarily identical to ours, it's easy to fall into misunderstandings harmful to the design of a complex product.

The browsing system is able to exploit knowledge extracted from textual material in several ways:

- expanding the search by exploiting the search for MWE and then offering suggestions to research incomplete or vague. For example, by proposing the query "Pontormo" are returned as results also the occurrences of "Jacopo da Prontormo", "Jacopo Carucci" and "Pontormo";
- improving the correlation process among documents identified, by using specific linguistic resources;
- improving the ranking of results in the classification of answers;
- better organizing the display of the results. The system shows in reply at the queries a variety of contexts. In these results are highlighted all relevant entities or "micro-semantics".

An example of browsing in WeBasCH shows that some material constitutes examples of grey literature. For example, the search of "premio" proposes the site of Aliano, particularly the web-page where planned activities for the year 2008 are displayed. This material would be lost over time, even though it provides important information about the natural and cultural heritage of a region.
In our approach, the creation of linguistic resources is designed to the development of navigation and information retrieval systems, that are able to exploit them. These tools capture, organize, classify and distribute the information in according to the desired objectives. In an open domain, as the Cultural Heritage, information can rarely be classified just with hierarchical criteria. An approach based on principles of "semantic similarity" is more efficient. This approach allows to link information crosswise, seemingly belonging to different categories, but that match to the same informative need.
The experience in the treatment of large amount of data has not only allowed the refinement of extraction tools for semantically relevant information, but also the creation of terminological resources toolkit.

The more a text is "enriched" with annotations, the better it can be processed by tools for analysis, categorization, browsing and Information Retrieval. The system is not limited to the identification and classification of entities; it also identifies the particular relations between the entities involved.

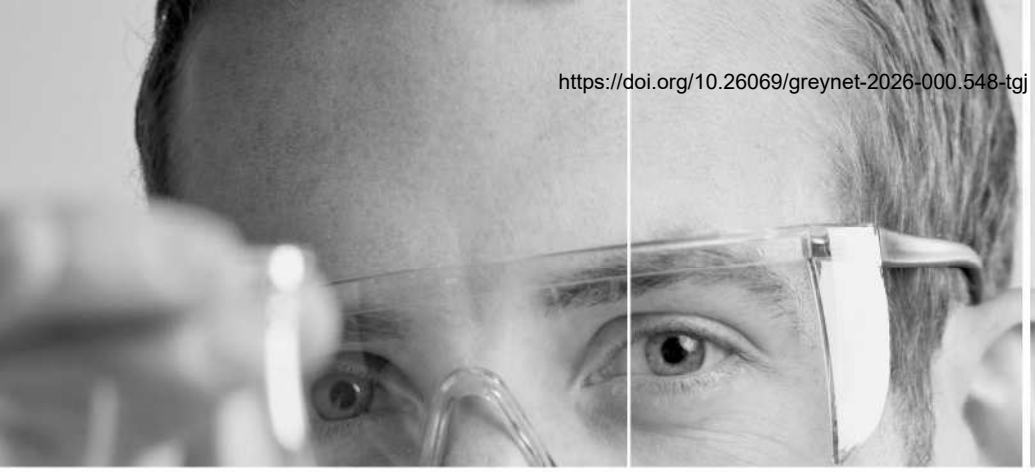
Conclusion

Our work proposes to overcome the traditional categorization systems and their rigidity, by means of a set of open and adaptive terms classes, which can guide the end user in refining of his/her search.
Such knowledge systems are valuable support on the one hand to networks of e-participation and e-government, on the other hand they offer more information and better performance.

References

1. Spadoni F., Tariffi F., Sassolini E., (2011). SMARTCITY: Innovative Technologies for customized and dynamic multimedia content production for Tourism applications. In: EVA 2011 Florence Electronic Imaging and the Visual Arts. (Firenze, 4-5-6 may 2011). Proceedings, Cappellini Vito (ed.). Pitagora Editrice Bologna, 130 - 135.
2. Granieri, G., Et Al., (2009) Linguaggi digitali per il turismo. Edizioni Apogeo, November 2006.
3. Picchi, E. Et Al. (2009). “Text Power”: tools for Cultural Heritage. In Proceedings of in 4th International Congress on “Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin”, IMC (CNR) Rome, Italy, pp. 277--278.
4. Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex ‘94, Amsterdam, The Netherlands.
5. Picchi E., Et Al., (2004). Linguistic Miner. An Italian Linguistic Knowledge System. In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 May 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.), 1811 – 1814.

https://doi.org/10.26069/greynet-2026-000.548-tgj



Slovak Centre of Scientific and Technical Information **SCSTI**

Achieve your goals with us

INFORMATION SUPPORT OF SLOVAK SCIENCE

SCIENTIFIC LIBRARY AND INFORMATION SERVICES

- technology and selected areas of natural and economic sciences
- electronic information sources and remote access
- depository library of OECD, EBRD and WIPO

SUPPORT IN MANAGEMENT AND EVALUATION OF SCIENCE

- Central Registry of Publication Activities
- Central Registry of Art Works and Performance
- Central Registry of Theses and Dissertations and Antiplagiarism system
- Central information portal for research, development and innovation - CIP RDI >>>
- Slovak Current Research Information System

SUPPORT OF TECHNOLOGY TRANSFER

- Technology Transfer Centre at SCSTI
- PATLIB centre

POPULARISATION OF SCIENCE AND TECHNOLOGY

- National Centre for Popularisation of Science and Technology in Society

IMPLEMENTATION OF PROJECTS

- National Information System Promoting Research and Development in Slovakia – Access to electronic information resources - NISPEZ
- Infrastructure for Research and Development - the Data Centre for Research and Development - DC VaV
- National Infrastructure for Supporting Technology Transfer in Slovakia - NITT SK
- Fostering Continuous Research and Technology Application - FORT
- Boosting innovation through capacity building and networking of science centres in the SEE region - SEE Science

www.cvtisr.sk
Lamačská cesta 8/A, Bratislava

Centralised National Corpus of Electronic Theses and Dissertations *

Július Kravjar and Marta Dušková (Slovak Republic)

Abstract

ETDs are a significant source of grey literature and not only for the academic community. Slovakia has made a big step forward by implementation of the Centralised National Corpus of Electronic Theses and Dissertations in 2010. The national ETD corpus consists of bachelor's, master's, dissertation and habilitation theses. This implementation was coupled with the concurrent implementation of the National Plagiarism Detection System (aka the originality check or the anti-plagiarism system). Both systems have to be used by all higher education institutions operating under the Slovak legal order. The new theses and dissertations incoming into the national ETD corpus are compared to this corpus and to the selected internet resources. The higher education institutions pay no fee for the service, the system acquisition costs were covered by the Ministry of Education, Science, Research and Sport, and the operating costs are also paid by the Ministry. The formation of both systems and the first two years of its existence is analyzed.

The first signs of activities towards ETD in Slovakia were recorded on the threshold of the Millennium. March 2004 was to become a significant milestone: sixteen academic libraries of twelve Slovak universities decided to solve the ETD.SK project: "Building Digital Academic Libraries - Collecting and Providing Access to Full Texts of Slovak University Publications". The ETD.SK project marked the beginnings of cooperation on a national level in this area, with the effort to follow up international ETD activities. Unfortunately, the project was not sufficiently implemented due to the lack of financial and personnel resources, but mainly because of the lack of legislative support.

The ICT and internet penetration, low copyright awareness and the rapid growth in the number of higher education institutions and students in our country contributed to the expansion of plagiarism. There was also an inherent lack of systemic action that would act as a barrier for its future growth. The establishment of the nationwide electronic theses and dissertation repository and their originality check was considered as a perspective solution.

A significant step in this matter was made in 2008: the Ministry of Education decided to implement a comprehensive nationwide solution for the collection and processing of theses and dissertations produced at Slovak higher education institutions. The goal: creation of the national theses and dissertations repository, increase in the quality of theses by their originality check, copyright and intellectual rights protection. In 2009, the Higher Education Act was amended and the most relevant change was this: Before the defence of the thesis, the higher education institution forwards the thesis in the electronic form to the Central Repository of Theses and Dissertations (CRTD) and the originality check is performed.

During 2009, the CRTD was built and the whole system with the originality check became reality at the end of April 2010. The existence of such a system has a preventive effect, and not just in the student community. CRTD is now a publicly-available source of grey literature.

A yearly increase in CRTD is about 80 thousand items of bachelor's, master's, dissertation and habilitation theses.

The paper analyzes the creation and two years' operation of the national corpus of bachelor, master, diploma, dissertation and habilitation theses of Slovak higher education institutions and the follow-up plagiarism detection system. The national corpus is called The Central Repository of Theses and Dissertations (CRTD). Each thesis has to be entered in CRTD before defence and it is then checked for plagiarism.

Creating Collections of Higher Education Theses: First Steps

"Creating digital collections of own academic production and making them available as full text digital documents, accessible in the computer network, has been gradually taken up by Slovak academic libraries in the last ten years. Projects aimed at the electronic processing of publications of university employees and at the electronic processing of theses and dissertations were the pilot projects in this area. Academic libraries implemented the first projects already at the beginning of the millennium.

* First published in the GL14 Conference Proceedings, February 2013.

March 2004 was a significant milestone in the area of Slovak academic digital libraries when 12 Slovak Universities presented the central development IT project "Building Digital Academic Libraries – Collecting and Providing Access to Full Texts of Slovak University Publications (ETD.SK)". ETD is the internationally used abbreviation for the university graduation theses in the electronic form, i.e. Electronic Theses and Dissertations. ETD.SK Project started the cooperation in this area at the national level with the effort to continue in international activities in the ETD area. The Project involved the specification of the organisational, technical and technological requirements for ETD collection, digitalisation workplaces were created in individual libraries and two hardware storage sites were built (in the cities of Košice and Prešov). Unfortunately, the Project was not implemented sufficiently in all libraries, mainly for insufficient financial and staff coverage; however, mainly for insufficient legislative support." (Haľko, 2011)
A positive effect was achieved as the universities started to make theses and dissertations accessible through online catalogues of academic libraries available on the Internet.

Plagiarism
Plagiarism existed in the past and will probably continue to exist in the future. In the Ancient Rome the word plagiarist (*plagiarius*) designated a kidnapper, mainly kidnapping children or slaves. Plagiarism action (*plagium, crimen plagii*) also meant offering refuge to an escaping strange slave.

The content of the word has been significantly modified and is kept only in a figurative meaning. Today "kidnapping" is understood as a theft of ideas, opinions or other intangible property that are illegally "appropriated" and presented to be original, authentic. Plagiarism can be defined as use of original ideas and creative formulations of another person with the aim to present them as one's own ideas or formulations. (Szattler, 2007)

In the absence of suitable instruments for the prevention and suppression of plagiarism, this phenomenon may overgrow to unacceptable dimensions. It would probably not be possible to fully eliminate all kinds of plagiarism in the future; however, it is necessary to build barriers where possible. We must not ignore or tolerate it.

The first higher education institution in Slovakia – the "early bird" – started using the system to reveal plagiarism in 2001 – it was a lonely runner during long seven years.

Background Situation
Slovakia with its 5.4 million inhabitants is confronted with plagiarism of higher education theses and dissertations like other countries. Plagiarism was growing at the higher education scene before 2010. There were no systemic measures preventing its growth. Rapidly increasing number of higher education institutions and students, growing Internet penetration, low understanding of copyright and intellectual property rights – this helped the growth of plagiarism. If we look at 1989 (year of Velvet Revolution[1]) as a basis – and compare it with 2011, the number of schools increased three times to 39 and the number of students increased four times to 250 000. In 1989, Internet penetration was zero; it reached 74.3% in 2010 and 79.2% in 2011.

How does the Internet help plagiarism? In two ways. The Internet offers free papers, final theses, dissertations and also expert literature. Insensitive use of the "copy and paste" functions with no adequate quotations compound in the text signed by the "author" – this is a typical example of plagiarism. Offers related to the preparation "supporting materials" for various types of theses can be found on the Internet with little effort. There are web pages offering the preparation of a wide range of theses (seminar, graduate, bachelor, diploma, dissertation, MBA etc.) covering markets of more than ten countries, and there are also web pages aimed only at local markets or markets with similar language or history. It is necessary to realise that "publication of another person's work as one's own is plagiarism". (SME.SK, Adamová, 2012)

Order and payment for the prepared bachelor, diploma or dissertation theses is not only the Slovak speciality, this is a global problem. If it is found before the graduation that the thesis was ordered, such student may be dismissed from his studies prematurely. If this is found after the graduation, nothing happens. Our Higher Education Act does not define the removal of a university degree. The degree will remain in the hands of the owner (Hospodárske noviny, 2012). It seems that this will change in the near future. The removal of fraudulently acquired academic degree should be defined in the Criminal Code as informed by the daily news in August 2012 (TASR, 2012).

The study of Králiková "Introduction of Rules of Academic Ethics at Slovak Higher Education Institutions" summarises the state of academic ethics at Slovak higher education institutions and shows that the

majority of inquired pedagogical workers directly experienced fraud of students. The most discussed topic in the media was the plagiarism of students, mainly the ways in which the students cheat and the instruments used by higher education institutions to eliminate plagiarism. Other themes included plagiarism of pedagogical workers or persons in high public positions. The absence of a wide discussion on academic ethics has also other consequences. One of them is that the academic environment members and the general public do not understand the importance of academic ethics and they are also less sensitive to a breach of it." (Králiková, 2009)

Issues related to the collection and processing of higher education institutions theses in electronic form and issues of plagiarism were often repeated in discussions in the academic area; however, with no significant progress. Seeds of future changes were seeded at the meeting of the Slovak Rectors' Conference in September 2006 (Slovak Rectors' Conference, 2006) when two documents on academic ethics were approved. These nationwide documents deal with ethics of pedagogical workers and students. However, the measures proposed to prevent plagiarism did not come to life (Králiková, 2009). In February 2008, the Slovak Rectors' Conference returned to the problems of plagiarism and asked The Ministry of Education, Science, Research and Sport of the Slovak Republic to coordinate activities connected with the acquisition of the plagiarism detection system. Higher education institutions were recommended to punish plagiarism and to create electronic archives of theses (Slovak Rectors Conference, 2009).

Although the plagiarism and the need to fight against it were being discussed a lot, the final decision was adopted by The Ministry of Education, Science, Research and Sport of the Slovak Republic in 2008 (in 2008, only two higher education institutions used the plagiarism detection system) for all higher education institutions operating under the Slovak legal order:

- To establish The Central Repository of Theses and Dissertations (CRTD), higher education institutions have to send any thesis or dissertation to CRTD;
- Any thesis and dissertation with the name and surname of the author and higher education institutions will be stored in CRTD for a determined period of time from the day of registration;
- Theses and dissertations must be checked for originality before defence, as proved by the originality check protocol;
- The originality check protocol is the result of the comparison with all the theses and dissertations in CRTD and to Internet sources and other available electronic sources;
- Higher education institutions' local repositories of theses and dissertations are the CRTD's communication partners.

The defined tasks were aimed to increase the quality of higher education institutions studies and also:

- Copyright and intellectual property rights protection, its better understanding;
- Theses and dissertations quality improved due to the originality check;
- The creation of The Central Repository of Theses and Dissertations;
- The creation of barriers to growing plagiarism.

Establishment of the Central Repository of Theses and Dissertations

In 2009, the Higher Education Act was amended and the most important modifications were:

1. The Ministry of Education, Science, Research and Sport of the Slovak Republic will administer the Central Repository of Theses and Dissertations
2. Before any person is allowed to defend his/her thesis/dissertation, the higher education institution sends his/her thesis/dissertation to the Central Repository of Theses and Dissertations in electronic form and undergoes the originality check.
3. Thesis/dissertation will be stored in the Central Repository of Theses and Dissertations with the name and surname of the author and the name of higher education institution for the period of 70 years from the day of registration. (Zbierka zákonov (Collection of Acts), 2009)

Theses and dissertations registered in the Central Repository of Theses and Dissertations after 31 August 2011 are publicly available – this was defined in the amendment to the Higher Education Act. (Zbierka zákonov, 2011)

The Central Repository of Theses and Dissertations of the Slovak higher education institutions and the Plagiarism Detection system became real at the end of April 2010. Higher education institutions operating under the Slovak legal order are obliged to use both systems. Approximately 80 000 theses are registered in the Central Repository of Theses and Dissertations yearly. As of 31 October 2012, there were 223,757 theses registered.

What did the implementation of the Central Repository of Theses and Dissertations and the plagiarism detection system (the originality check) bring?

Before the implementation of the project, it was necessary to consider and address:

- The centralised national corpus of electronic theses and dissertations;
- The growing plagiarism of theses and dissertations;
- The occasional efforts of higher education institutions to check originality of theses and dissertations;
- The absence of system instruments to fight against plagiarism at the national level;
- The public access to theses and dissertations from one spot;
- The increased understanding of copyright and intellectual property rights.

The published information on the prepared centralised national corpus of electronic theses and dissertations and the originality check brought preventive effect even before the implementation. Other effects of the implemented project were the following:

- The principal breakthrough in the fight against plagiarism in Slovakia, obligatory use of the Central Repository of Theses and Dissertations and the originality check (with the Central Repository of Theses and Dissertations, Internet sources and other electronic sources).
- The existence and real operation of the instrument for the protection of copyright and suppression of plagiarism.
- Unified methods for collecting theses and dissertations (creating centralised repository for all higher education institutions) and unified system for the originality check (Plagiarism Detection System).
- Automated collection of theses and dissertations, originality checks and distribution of originality check protocols.
- The existence of CRTD (as one of the sources of grey literature) and the system to detect plagiarism is preventive in the community of students and others. It increases the understanding of copyright and intellectual property rights at least in the academic area; it improves how students work with literature, Internet, quotations, and improves the quality of theses and dissertations.
- All higher education institutions in Slovakia operating under the legal order of the Slovak Republic are obliged to use the complex of the CRTD with the originality check at the national level. This applies to 35 higher education institutions of 39.
- All theses and dissertations are collected in the CRTD where they are stored for 70 years.
- The public may verify the suspected plagiarism on the web page where theses and dissertations are published: http://www.crzp.sk/crzpopac?fn=*searchform.
- The originality check protocol does not confirm that a thesis/dissertation is original or a plagiarism. The protocol is the basis for the examination committee's decision, it helps – it informs about documents a supervisor or an opponent might have overlooked. The originality check protocol identifies the parts of the text of the presented thesis/dissertation identical with the parts of texts deposited in CRTD and with Internet sources. The originality check protocol is available to the examination committee for evaluation and it is a part of the final (national) exam records.
- Higher education institutions do not pay to use these systems; the procurement costs were covered by the Ministry of Education, Science, Research and Sport of the Slovak Republic and the operation costs are also paid by the Ministry.

In Conclusion

There are large reserves in the fight against plagiarism in the education of the young generation. Plagiarism should be prevented and the process of education to the non-cheating culture must start gradually and adequately from the earliest age, from the pre-school education level. (Skalka, et al., 2009).

A correctly oriented and properly timed educational process and the implementation of advanced technologies have high potential in the fight against plagiarism. Technologies, however, are not a panacea. Very important and irreplaceable is the role of education – from the beginning of the educational process – in close connection with the prevention and plagiarism uncovering with clearly defined rules and penalties, and with mutual interaction of all these parts. (Kravjar, 2011).

The implementation of the CRTD together with the originality check at the national level in everyday practice is a unique solution in Europe and probably in the world. The system has high potential to be implemented in many areas. The originality check can be done where a written work is the outcome.

The following may be considered:

- Seminar and other works at higher education institutions;
- Research reports;
- Applications for projects, grants, their outcomes;
- Final works for increasing qualification of pedagogical and other professions;
- Secondary school works within the curriculum;
- Secondary school works beyond the curriculum;
- General publication activities (the establishment of relationships with publishers);
- The repositories of grey literature.

The system is still being developed. Last year, the system supplier (a company called SVOP, Ltd, http://svop.sk/en/Default.aspx) won the first prize at the international competition of anti-plagiarism systems in Amsterdam "PAN 2011 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse" held in conjunction with the "CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation", where their algorithm detecting plagiarism was the best in all four indicators (plagiarism detection, recall, precision, granularity). It is necessary to remind that the competition corpus did not include Slovak texts, only English, German and Spanish texts, and to point out that the system was also able to uncover the so-called "translated plagiarism", and that the detection is invariant against a change of word order, against the occurrence of changed words, against omission or additions of words in the text in a suspicious document.

The plagiarism detection system received a significant award at the international conference ITAPA 2011 in Bratislava (Information Technology and Public Administration). In the category New Services, the nationwide plagiarism detection system and CRTD ranked second.

This paper is a reworked and updated version of the conference paper Barrier to thriving plagiarism (Kravjar, 2012).

References

HAĽKO, P. (2011). *Elektronický zber záverečných prác na Prešovskej univerzite v Prešove*. [ONLINE] Available at: http://www.pulib.sk/elpub2/PU/Uninfos2011/data/ P11_Halko.pdf. [Accessed 26.6.2012].

KRÁLIKOVÁ, R. (2009). *Zavádzanie pravidiel akademickej etiky na slovenských vysokých školách.* [ONLINE] Available at: http://www.governance.sk/assets/files/publikacie/akademicka_etika .pdf. [Accessed 21 March 2012].

KRAVJAR, J. (2012). *Barrier to thriving plagiarism*. [ONLINE] Available at: http://www. plagiaris madvice.org/documents/conference2012/finalpapers/Kravjar_fullpaper.pdf. [Last Accessed 22.8.2012].

KRAVJAR, J. (2011). *Antiplagiátorský systém.* [ONLINE] Available at: http://www.inforum.cz/pdf/2011/kravjar-julius.pdf. [Accessed 23 March 2012].

SKALKA, J. et al. (2009). *Prevencia a odhaľovanie plagiátorstva.* [ONLINE] Available at: http://www.crzp.sk/dokumenty/prevencia_odhalovanie_plagiatorstva.pdf. [Accessed 21 March 2012].

SZATTLER, Eduard. *Právne a morálne aspekty plagiátorstva. Duševné vlastníctvo : Revue pre teóriu a prax v oblasti duševného vlastníctva* [online]. 2007, 1, [cit. 2011-04-06]. Dostupné z WWW: <http://www.solain.org/clanky/20070325192403.pdf>.

SME.SK (e.g. 2011). *Veľký biznis: Diplomovka na kľúč v akcii za pár stoviek eur* . [ONLINE] Available at: http://www.sme.sk/c/6329494/velky-biznis-diplomovka-na-kluc-v-akcii-za-par-stoviek-eur.html. [Accessed 28.6.2012].

HOSPODÁRSKE NOVINY (2012). *Napíšem vám diplomovku. Stačí zaplatiť.* [ONLINE] Available at: http://style.hnonline.sk/vikend/c1-55146090-napisem-vam-diplomovku-staci-zaplatit. [Accessed 23 March 2012]

SCHÖPFEL, J. (2010). *Towards a Prague Definition of Grey Literature*. [ONLINE] Available at: http://archivesic.ccsd.cnrs.fr/docs/00/58/15/70/PDF/GL_12_Schopfel_v5.2.pdf. [Accessed 22.8.2012].

TASR.SK (2012). *Podvodne získaný titul bude trestným činom* . [ONLINE] Available at: http://www.sme.sk/c/6511466/podvodne-ziskany-titul-bude-trestnym-cinom.html. [Accessed 30.8.2012].

ZBIERKA ZÁKONOV (2009). *Zbierka zákonov č. 496/2009.* [ONLINE] Available at: http://www.zbierka.sk/sk/predpisy/496-2009-z-z.p-33272.pdf. [Accessed 20 March 2012].

ZBIERKA ZÁKONOV (2011). *Zbierka zákonov č. 6/2011.* [ONLINE] Available at: http://www.zbierka.sk/sk/predpisy/6-2011-z-z.p-33957.pdf. [Accessed 20 March 2012].

[1] The Velvet Revolution (Czech: *sametová revoluce*) or Gentle Revolution (Slovak: *nežná revolúcia*) was a non-violent revolution in Czechoslovakia that took place from November 17 to December 29, 1989. Dominated by student and other popular demonstrations against the one-party government of the Communist Party of Czechoslovakia, it saw to the collapse of the party's control of the country, and the subsequent conversion to capitalism (Wikipedia).

https://doi.org/10.26069/greynet-2026-000.548-tgj

An Environment Supporting the Production of Live Research Objects*

Massimiliano Assante, Leonardo Candela, and Pasquale Pagano (Italy)

Abstract

Modern science communication requires innovative environment and means for providing stakeholders with scientific outcomes. Research objects are emerging as replacements of traditional "documents" in scientific communication. These objects are multi-media and multi-part objects that aggregate all the "pieces" that contribute to a research result. Supporting these objects has gone beyond the capacity of traditional technological approaches based on locally specialized data management facilities. In this article we present an environment for producing "live research objects" by exploiting the capabilities offered by a Data Infrastructure. Such environment includes: (i) a workspace where users can organize and share with their co-workers very different items in a file-system-like environment; (ii) an editing framework where users can define the structure of a live research object and compile objects that comply with one of the defined templates; and (iii) a workflow engine where users can define the workflow governing the production of a live research object by specifying the phases and the relative responsible actors(s).

1. Introduction

Scientific research is rapidly evolving in all fields, it is multidisciplinary, networked and driven by new patterns, e.g. data-intensive sciences [1]. In this complex scenario scientific communication must go well beyond traditional scholarly communication. Specifically, it requires accessing all the elements exploited and developed during the scientific workflow to achieve a result, e.g. datasets, analysis tools, and methods [2]. This wide corpus of primarily grey elements are at the moment mostly unavailable and, even when they are available, they are not linked to the scientific result. This makes difficult to completely understand the result and validate it.

To overcome this limitations many initiatives have been proposed in the literature as replacement of traditional "documents" in scientific communication, such as *Executable Papers* [16,17], *Enhanced Publications* [6,15], *Living Reports* [8,7]. Some of them were sponsored by major scientific publishers [12].

Lately, *Research Objects* [10] are emerging as an abstraction for communicating, sharing and reusing research results. These are multi-media and multi-part objects that aggregate all the "pieces" that contribute to a research result. Such elements, which may range from binary files to compound objects including maps, time series, and tabular data, are generally structured according to well-established templates and produced according to user-defined workflows. We extended the Research Object definition and included facilities to make some of these "pieces" contributing to a research result directly embedded in the document. The idea behind a Live Research Object is depicted in figure 1.

However, supporting Research Objects has gone beyond the capacity of traditional technological approaches based on locally specialized data management facilities. This paper discusses how an infrastructure-oriented approach aimed at promoting *sharing* and *re-use* of resources (including data, services and computational and storage resources) is an effective approach for producing Live Research Objects and poses the accent on the user-oriented facilities supporting the collaborative production of such research products.

* First published in the GL14 Conference Proceedings, February 2013.

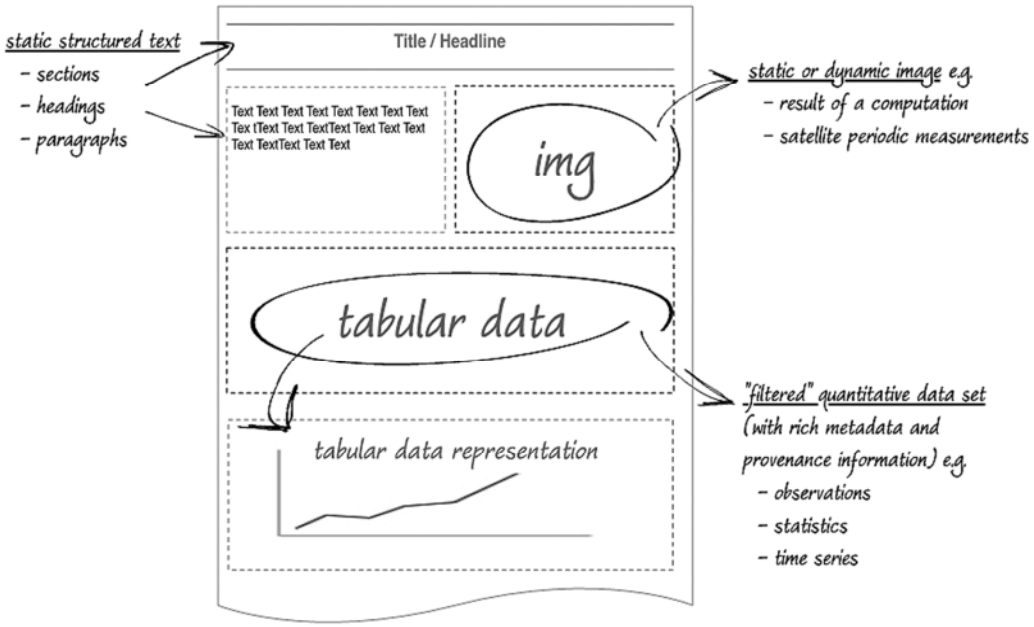


Fig. 1. The Idea behind a Live Research Object

The production environment includes: (*i*) a *workspace* where users can organize and share very different items (from binary files to compound objects) in a file-system-like environment; (*ii*) an *editing framework* where users can define the structure of a live research object (a template indicating sections, layout, active elements) and compile objects that comply with one of the defined templates by entering content or taking it from the workspace via *drag and drop*; and (*iii*) a *workflow engine* where users can define the workflow governing the production of a live research object by specifying the phases and the relative responsible actors(s).

Accordingly, the article is structured as follows. Section 2 describes the requirements and the main design decisions driving the development of the proposed approach. Section 3 presents the environment developed to offer Live Research Objects production. Finally, Section 0 concludes the article and summarizes its results.

2. Motivations and Design Philosophy

We feel that the classic notion of documents viewed as static entities has to be abandoned. This notion should be moving towards one where documents are constantly evolving, by enriching them with components that yield seamless data access to contents, user cooperation, collaboration, interaction, and personalization.

However, this notion cannot evolve if not supported by progresses in the way documents are created, produced and made available too. In fact, thanks to the new technologies, scientists, researchers or experts in a field, can produce novel research outputs based on new resources (in terms of hardware, services and content) that were not available in the past.

Despite that, this production can still require a lot of work due to *(i)* the complexity of interfacing with different sources and tools, and *(ii)* the people involved in the task that may need coordination, concurrent and rule-based access to the same research object or part of it. In particular, for the research object instance case the resulting work may also not meet the requirements of its consumer, *i.e.*, the reader, since it could present a picture of the subject at the time of its production and not at the time in which the information produced is accessed and used.

The latter point illustrates the potential for new scientific advances. For example, the Food and Agriculture Organization of the United Nations (FAO) exploits its rich information sources, ranging from raw data sets to graphs and map archives, to periodically prepare reports on the status of the agriculture and fishery, per country. This activity often:

- make necessary to have access to a wealth of textual, graphical and tabular information located in several (external) data sources;
- requires several people working together in the collection, collation, drafting, and reviewing;
- demands for “freshness” of the information reported.

To overcome these problems we decided to experiment the construction of a workflow[i]-driven Live Research Objects production environment. This environment *(i)* exploits the facilities provided by an

underlying *Hybrid Data Infrastructure* [18] for accessing and exploiting the entire spectrum of resources needed to achieve a research result including data, services and computing resources "as-a-Service", *(ii)* uses a component-oriented flexible document model for the representation of the research objects, *(iii)* benefits from a workflow driven mechanism to ensure concurrent and rule-based access to its users, and *(iv)* is accessible through a thin client (namely a web browser).

The solution for the described approach can be logically divided in two main modules: one module realizes an environment for producing (namely defining and editing) innovative research objects, one module realizes an environment driving users (namely assigning actions and notifying the right actor(s) when needed) while producing such research objects.

The first module is delegated to solve the issues related to the complexity of interfacing with different sources and tools, providing users with a familiar and easy-to-use environment to produce research objects. The following design decisions have been taken for the implementation of the proposed approach:

- The management of different data sources should be made transparent to the end-users (by the hybrid data infrastructure) and promote the seamless access and sharing of a rich array of information objects;
- The research objects production environment should enable a WYSIWYG[ii] editor (similar to Google docs) to give users an immediate feeling on how the actual research object will look like;
- The production environment should enable the production of research objects sharing a common structure, when needed. Thus it supports a production based on two phases: *(i) template definition*, to define a basic research object structure (including sections and layout) to be adopted by research objects expected to be compliant with such a structure, and *(ii) reporting*, to produce the actual research object in compliancy with one of the defined templates;
- Supported Research Objects should enable the definition of living elements, i.e. Research Object elements that are potentially willing to evolve whenever an user accessed the object;
- The produced research object should be available in different exporting formats, including OpenXML, PDF, and HTML.
- The second module of the solution instead is in charge of coupling the research object with a workflow driven mechanism and ensuring through a series of steps, rule-based access to the same instance. For the implementation of this module the following design decisions have been taken:
- support for visual representations of the workflow, through workflow diagrams outlining the whole process;
- support for workflow states;
- support for manual and automatic (policy-based) routing;
- support for routing to individuals and to groups (set of users having the same role).

Accordingly, the next section presents the environment developed to offer workflow driven Live Research Objects production facilities.

3. The gCube Live Research Objects Environment

gCube[iii] is software system enabling the building and operation of an Hybrid Data Infrastructure. In a nutshell, it offers a number of mediators for interfacing with (data) providers and a number of services for data management over a rich array of data types. The gCube Live Research Objects Environment is one of these data management services and it implements the environment envisaged in the previous sections to realize Live Research Objects. It is made of three main components: (i) the *Virtual Workspace*, that is a virtual file system promoting the sharing of a rich array of information objects, (ii) the *Research Object Editing Environment*, that is a graphical editor and renderer of a Live Research Object, and (iii) the *Research Object Workflow Manager*, that allows to create and associate workflows to a given Research Object and to manage and control its status.

3.1 Virtual Workspace

The Virtual Workspace (WS) is the core element of the cooperation environment. It is conceived to resemble a classical folder-based file system any user is familiar with. As a consequence, the operations it supports on the items are the expected ones, namely items creation, deletion and their organization in folders and subfolders. Thus every single user is free to organize its items.

However, the real added value of this file-system-like environment is represented by the types of items it can manage in a seamless way. They range from binary files to information objects representing tabular data, species distribution maps, and time series. Every item in the workspace is equipped with a rich metadata including bibliographic information like title and creator as well as lineage data.

Another distinguishing feature is represented by the sharing that is fundamental since users need to work collaboratively on the same research object and rely on common research materials. Sharing can be performed per single item as well as per folder and it is invite-based. Any item or folder and, in turn, its content can be shared with other users. The users involved in this sharing are alerted by a notification mechanism in charge of delivering the related invite.

The end-user interface

A web application has been developed to offer users the possibility of viewing and managing their research object instances and as well as any other information object. This web application offers a remote file system view similar to any Operating System over the content available within the data infrastructure along with files management facilities such as copy, delete and upload.

3.2 Research Object Editing Environment

Document templates motivate their uptake by easing the strain of repetitive manual work. It is quite common that a document's structure and style is maintained through time and used for multiple document instances. This is the reason why the Research Object Editing Environment, that is a web application developed by using the Google Webtool Kit framework [13], is logically divided in two phases: the Template Definition and the Research Object Editor. The first one is needed to define research object templates that, during this stage, are dynamically and statically completed. The second instead is capable of loading these templates to produce actual research objects by filling out their dynamic parts. In the following we explain our design approach from a functional description point of view together with actual examples for both phases.

Template definition

During this phase, document templates can be created through user interfaces by exploiting an extension of the Document Editor web application, called Template Creator.

The Template Creator adopts a component oriented approach for templates composition and supports several component types such as structured and styled text areas (title, headings, body), images, tables, table of contents (ToC), bibliography, page breaks. Template components are divided in two classes: *static* and *dynamic*. A static component of a template is, as the name suggests, a part that is not meant to change across diverse research objects sharing the same template, such as filled-in texts or images. A dynamic component instead is meant to be completed in the second phase and, in turn, can belong to the following two categories:

- *dynamic text*: empty text areas, including headers and empty tables, fall into this category. An example would be having the same template for a set of research objects aiming at representing a study on a species. One would change species' specific data while the structure and the presentation would be uniform for all of them.
- *dynamic spot*: empty rectangular spots designed to host an image, a table, a diagram or a chart, to be instantiated using a given data source or data set. This is a key category as during the second phase these spots become configurable, providing users with the possibility of entering configuration parameters. Examples vary depending on the "hosting type" and will be explicated in the following.

The template components belonging to a template can be grouped into sections. It is a human-friendly interface assuming the form of an HTML page. This application provides users with toolbars and toolboxes from where they have the possibility of adding or removing template components, adding or removing sections, formatting text and saving their work into the Virtual Workspace.

Research Object Editor: The Document templates, created through the Template Creator extension, can be loaded by accessing the Virtual Workspace from within the Reporting Editor through user interfaces.

The Research Object Editor in this phase offers the possibility to complete or instantiate the dynamic components that might be present in a template. Depending on their type, this action can be performed in two ways: *(i)* by typing in, for components belonging to the *dynamic text category* (formatting text as one would in a word processor by using a formatting bar) or *(ii)* by providing a configuration for components belonging to the dynamic spot category. This configuration can vary depending on the *dynamic spot* type. For instance, users would specify which column and which row intervals to show for *tables* or, what data to be set on x and y axis for *charts* and so on.

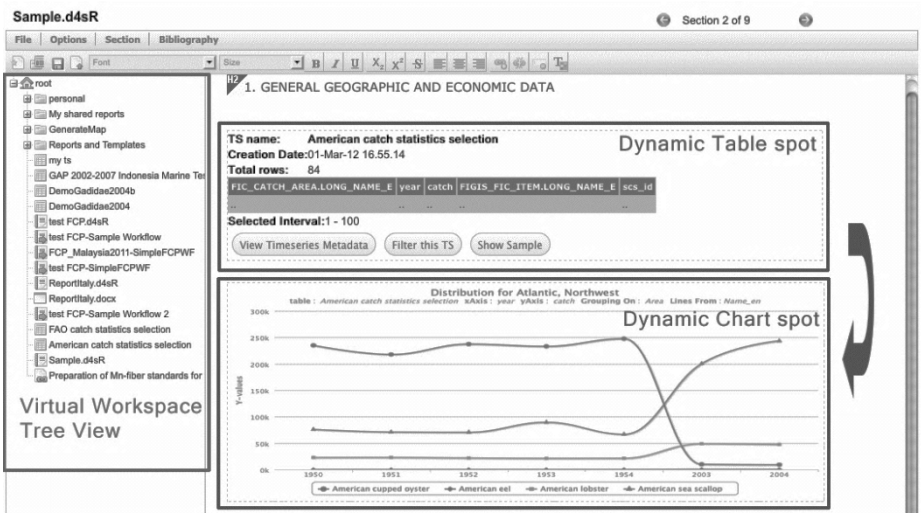


Fig. 1. A partial view of a template being completed in the Research Object Editor

Figure 1 illustrates the Research Object Editor user interface, a tree view of the Virtual Workspace is presented next to the *working area*. This tree view is needed to connect dynamic spots to the actual data they need to display. This connection is made explicit through *Drag and Drop* operations (by dragging the desired item over the desired spot).

Figure 1 also shows an example of template section being completed next to the tree view. The section is composed by three components, a static heading component on the top and two dynamic spots below, the first of type Table and the second of type Chart. In the example, both spots have been connected to the same *quantitative dataset*[iv], that is represented as an item in the user's Virtual Workspace, and describes a catch statistics *time series*[v] (reporting statistics on catches of marine species on a give geographic area, per year). The example shows the type Table being configured while the type Chart, that has been already, is actually displaying the related chart.

It is important to stress the fact that the *dynamic spot* component during this phase becomes *living, interactive* and capable of redisplaying itself depending on the parameters specified by the user.

Research Object instances can be successively exported into different formats by using functions provided by this editor, such as OpenXML (docx), PDF, HTML and saved into the user's personal Virtual Workspace.

3.3 Research Object Workflow Manager

The idea behind the Research Object Workflow Manager (ROW) is to work collaboratively to the creation of Research Objects. During this phase the Research Object is being created and is still incomplete. It is initiated by an actor, generally identified by a role, and then passed to other people involved in the realization. In the following, to distinguish this phase we will use the term Report instead of Research Object.

For example, one author is asked to start drafting the Report, successively several iterations can be possible between authors and editors, before the Report is sent for review and authorization. To support such scenarios, ROW is equipped with a *workflow roles editor* that, as the name suggests, allows to distinct the users involved in the creation of the Report into categories *e.g.*, *author, editor, publisher.*

The ROW is also equipped with a *workflow templates editor*, a graphical editor and renderer of a workflow diagram. A *workflow template* (WT) is naturally represented as a digraph (directed graph) where vertices, *i.e.*, workflow steps, are connected by directed edges, or arcs. WTs support the possibility to apply *labels*, *i.e.*, *roles*, on edges to indicate the required role to perform the transition from one step to another and are always defined by an entry and an ending point, i.e., Start/End steps are mandatory in any WT. An example of a workflow template, loaded in this editor is depicted in Figure 2.

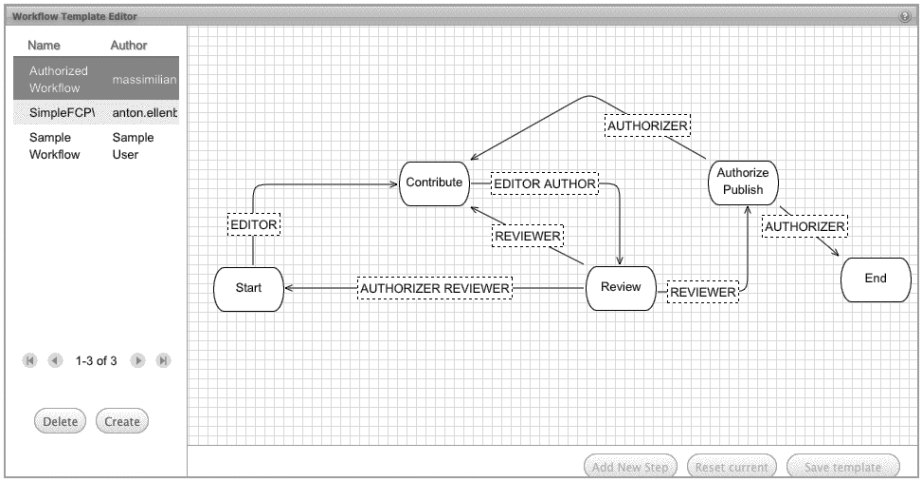


Fig. 2. A workflow template diagram loaded in the Workflow Templates Editor.

The presence of workflow templates is justified by the fact that reuse is highly desirable. Often one would like to reuse the same workflow diagram for multiple Reports, alternating the users involved but the process.

Roles and Templates are indeed complementary components for the ROW: they are required for the creation phase of a workflow document, where a Report's draft is linked to a Workflow Template, and for the functioning and monitoring of the reporting activity, *i.e.*, the step-by-step procedure needed to complete the job.

In particular, the creation phase of a new workflow document is performed by an entitled user and requires the three stages illustrated in Figure 3. In the first stage the user associates a Report draft to a Workflow Template: selects the Report to work with, created with the Research Object Editor by accessing the Virtual Workspace and couples it with an already available template (created previously through the workflow template editor). The second stage requires the user to specify, for each step, the associated roles and their relative permissions. It is possible to associate any Role to a given step, however Roles (labels) connected to a step's outgoing edges are mandatory and already present for permissions to be added. It is worth noting that the permissions attached to a role can vary from one step to another, *e.g.*, during the step "review" an author has read-only permission while during the step "contribute" an author has both read and update permissions. During the third stage instead the steps are not interested while roles, defined in the workflow report during the second stage are. In fact this is the part where these roles are linked to the actual users partaking in the workflow report production.

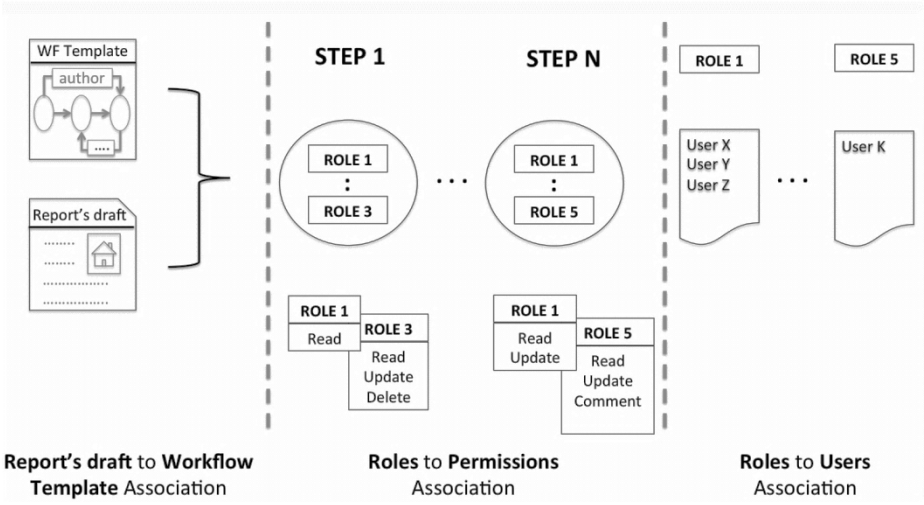


Fig. 3. The three stages needed for the creation of new documents workflows.

Regarding ROW *functioning*, manual routing and automatic routing are supported as by requirement of Section 2. The former is achieved by providing the workflow reports owner (WFO), *i.e.*, the user who created workflow reports through the three stages procedure previously described, the possibility to decide when and where performing step transitions, for the latter instead, the concept of forward action has been introduced. A *forward action* is a sort of green flag users make us of to indicate their work is completed for the current step (in which they are required to do a job). For instance, suppose there are three authors assigned to a step "contribute", each author completes his job and then performs a forward action towards the next step, *e.g.*, "review", ROW's logic recognizes it and consequently executes the transition (automatically) towards the expected step, *i.e.*, "review" for the sake of this example. It is worth noticing that the policy applied to automatic transitions is the one of executing it if, and only if, the totality of the actors involved in a job have forwarded. However, this can be changed to majority (instead of totality), by setting a parameter during the workflow report's creation phase.

Regarding ROW *monitoring*, the WFO can continuously monitor the users activities involved in a workflow. Clearly, a WFO can see and monitor only the workflow reports he owns. For each workflow report it is possible to *(i)* check the current status of the workflow, that is characterized by the current step and the forward actions performed until that time, *(ii)* check the workflow history, a chronological reversed list of user actions on the workflow.

The concurrent access, that is required since multiple users could work on the same report instance at the same time, is guaranteed by a locking mechanism provided by the ROW. When a user is editing a report instance a lock is acquired over it and no other user can access the report until the editing one finishes his work and commits the changes. The information a user has over any workflow report he needs to work with are the following: *name, current status, his role, date of creation, last action performed, grants (read, update etc.), whether the report is locked or not.*

4. Conclusion

Scientific research is nowadays multidisciplinary, networked and driven by new patterns, *e.g.*, data intensive sciences. This calls for innovative approaches and tools that are capable to cope with disciplines and tasks that can not be tackled by researches operating in isolation. Moreover, research products expected to be produced are richer than traditional research products, namely scholarly publications.

In this paper, an innovative environment for the production of live research objects was presented. In particular, a number of facilities supporting the whole lifecycle leading to the production of Live Research Objects was described. These facilities include: *(i)* a shared workspace resembling a classical file system and enabling users to store and organize any research artifact leading to a research product; *(ii)* an editor conceived to support the definition of research object structures, to promote the realization of research objects compliant with a given structure by implementing a number of user friendly features, *e.g.*, drag and drop of object constituents from the user workspace; and *(iii)* a workflow engine supporting the definition and operation of workflows driving the realization of a live research object in a collaborative and distributed approach.

Acknowledgments. The work reported has been partially supported by the D4Science-II project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and the iMarine project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644).

References

1. T. Hey, S. Tansley, and K. Tolle. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, 2009.
2. Boulton, G.; Campbell, P.; Collins, B.; Elias, P.; Hall, D. W.; Laurie, G.; O'Neill, O.; Rawlins, M.; Thornton, D. J.; Vallance, P. & Walport, M. Science as an Open Enterprise. *The Royal Society,* 2012
3. http://en.wikipedia.org/wiki/timeseries. TimeSeries Definition, 2012.
4. M. Altman and G. King. A Proposed Standard for the Scholarly Citation of Quantitative Data. 13(3/4), Apr. 2007.
5. T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni. Deploying general purpose virtual research environments for humanities research. Philosophical Transactions of the Royal Society A, 368:3813–3828, 2010.
6. OpenAIRE EU Project Website - What is an Enhanced Publication? http://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications 2012
7. L. Candela, F. Akal, H. Avancini, D. Castelli, L. Fusco, V. Guidetti, C. Langguth, A. Manzi, P. Pagano, H. Schuldt, M. Simi, M. Springmann, and L. Voicu. DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. International Journal on Digital Libraries, 7(1- 2):59–80, October 2007.
8. L. Candela, D. Castelli, P. Pagano, and M. Simi. From Heterogeneous Information Spaces to Virtual Documents. In E. A. Fox, E. J. Neuhold, P. Premsmit, and V. Wuwongse, editors, Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Lecture Notes in Computer Science, pages 11–22, Bangkok, Thailand, December 2005. Springer
9. G. Crane, A. Babeu, and D. Bamman. eScience and the humanities. International Journal on Digital Libraries, 7(1-2):117–122, October 2007.
10. Belhajjame K., Goble C., & De Roure D. Research object management: opportunities and challenges. Data Intensive Collaboration in Science and Engineering (DISCOSE) workshop, collocated with ACM CSCW 2012.
11. G. Crane, A. Babeu, and D. Bamman. eScience and the humanities. International Journal on Digital Libraries, 7(1-2):117–122, October 2007.
12. The Executable Paper Grand Challenge, Elsevier http://www.executablepapers.com/
13. Google Inc. Google Webtool Kit. http://developers.google.com/web-toolkit/
14. J. Lave and Wenger. Situated Learning: Legitimate Peripheral Participation. Cam, 1991.
15. S. Woutersen-Windhouwer, R. Brandsma, P. Verhaar, A. Hogenaar, M. Hoogerwerf, P. Doorenbosch, E. Durr, J. Ludwig, B. Schmidt, B. Sierman, "Enhanced Publications", edited by M. Vernooy-Gerritsen, SURF Foundation, Amsterdam University Press, 2009
16. Van Gorp, P. & Mazanek, S. SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science,* 2011*, 4*, 589-597
17. Nowakowski, P.; Ciepiela, E.; Harezlak, D.; Kocot, J.; Kasztelnik, M.; Bartynski, T.; Meizner, J.; Dyk, G. & Malawski, M. The Collage Authoring Environment. *Procedia Computer Science,* 2011*, 4*, 608-617
18. Candela, L.; Castelli, D. & Pagano, P. Managing Big Data through Hybrid Data Infrastructures. *ERCIM News,* 2012, 37-38

[i] Commonly agreed, workflow is the series of steps procedure taken to complete a given task or job.
[ii] WYSIWYG is an acronym for "what you see is what you get". A WYSIWYG editor is one that allows a user to see what the end result will look like while the document is being created.
[iii] www.gcube-system.org
[iv] A quantitative data set represents a systematic compilation of measurements intended to be machine readable. The measurements may be the intentional result of scientific research or information produced by governments or others for any purpose, so long as it is systematically organized and described [4].
[v] A time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals [1].

A Funder Repository of Heterogeneous Grey Literature Material with Advanced User Interface and Presentation Features *

Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos, Despina Hardouveli, Alexandra Roubani, Ioanna Sarantopoulou, Alexandros Soumplis, and Chrysostomos Nanakos (Greece)

1. Introduction

The present contribution concerns the development of a funder repository aiming at the dissemination, reuse and preservation of mainly grey literature material of diverse types. This material which was produced under the auspices of large scale (multi-billion Euros) funding programmes of the Hellenic Ministry of Education, co-financed by the European Union. The project involved the handling of a wide range of content, like studies, reports, educational material, videos, theses, material from a range of conferences/events. The project has been successfully completed and the system is publicly available since spring 2011 at http://repository.edulll.gr.

In this repository creation use case, technical enhancements were used to provide the means to organise, process and import to the repository a wide range of heterogeneous material. Special facilities for the support of these workflows was included, along with enhanced user viewing capabilities for metadata, PDF files and video content, providing an end user experience suitable for the repositoty's users, which include the educational community, researchers and scientists and the general public The metadata schema used is an application profile using elements for Qualified Dublin core, LOM and PREMIS. The repository has been implemented using the DSpace platform, with significant extensions developed specifically for this project. Futhermore a supplement external lightweight CRIS system has being developed in internal beta version, in order to appropriately represent semantically rich information.

The rest of the article at first explains the overall project scope and the followed methodology and approach, including important issues encountered and relevant decisions that had to be made. Implementations aspects, particularly regarding metadata schema and workflows are presented in Section 3. Advanced repository features are described in Section 4 and the paper concludes with summary, lessons learnt and ideas for future work.

2. Project scope, methodology and approach

Since 1994, the Hellenic Ministry of Education has been executing a series of large scale (at the range of billion of Euros) funding programmes, co-financed by the European Union through Structural Funds, under the names Operational Programme for Education and Initial Vocational Training (EPEAEK I & II, 1994-1999 and 2000-2006, respectively) and Operational Programme for Education and Lifelong Learning (2007-2013). The entity responsible for the management of the programmes is a Special Managing Authority within the Ministry (www.edulll.gr). The programmes finance a variety of activities related to all level of education (i.e. K-12, Higher Education and Lifelong Learning) including, among many others, restructuring and enhancement of educational programmes, production of educational material, scholarships and fellowships for doctoral students and post-doctoral researchers, collaborative research projects with the participation of Greek Higher Education Institutions, as well as a wide range of studies and reports, conferences and events, seminars, etc.

In autumn 2010, the respective Special Managing Authority assigned to the Hellenic National Documentation Centre the task of organizing the digital material that has been produced in the course of the aforementioned funded activities into a publicly available repository that would provide a single point of access to the programmes results, ensure wide dissemination and reuse of content through enforcement of interoperability standards and also address certain long-term preservation aspects.

The input material available from the Special Authority had the form of digital files, arranged in folders mainly according to project, however without any metadata attached to them. A considerable portion of these primary sources concerned administrative reporting during project execution and related files (e.g. administrative reports, detailed schedules of courses and seminars, participants' lists for seminars / events, etc.) It was decided to ignore this kind of material so that the repository includes only items that contain content of archival value. It is worth noting that during this first-phase project also a lot of the digital material produced during these programmes was not included, simply due to non-availability in digital form.

* First published in the GL14 Conference Proceedings, February 2013.

Processing of digital files before publication via the repository was a challenging issue due to the inherent heterogeneity of the material and the non-uniformity in formats, naming and structure. A considerable part of the overall project concerned this aspect. An important decision was to store and preserve both initial and processed digital files within the repository so that potential problems in the sometimes non-trivial processing workflow could be corrected a posteriori. Notably, providing perpetual persistent access to the the digital material and handling certain preservation issues was an important goal of the project; among the main reasons for that was the fact that a significant part (probably the majority) of the content concerned grey literature that would not be handled through official publishing and preservation channels.

Regarding metadata schema, a custom application profile was created, based on existing common international standards. This has been the result of an iterative process. The first phase consisted of a detailed mapping of the input digital material during which the most common types of items were identified and candidate fields per type were defined. This activity was necessary due to the diversity of the material and the limited available information initially on types and formats of content. It has been a tedious – despite assisted to some extent by automated tools – process of going through the directories/files and capturing types of items, formats and doing an initial distinction of content candidate for inclusion to the repository and administrative documents. The first phase produced a first version of the metadata schema that was to a large extent stable for the most common document types and enabled the beginning of the cataloguing process early in the project, followed by subsequent phases of refinement. Further details on the metadata schema are provided in paragraph 0.

Certain important cataloguing decisions were made early on. First, there was a considerable allocation of effort to subject cataloguing and appropriate user interfaces for browsing by subject (see paragraphs 0 and 0), since the diversity of the material would make it inadequate to rely only on navigation/browsing based on fields like keywords, type, project or collection. Secondly the idea of self-archiving was abandoned, at least for this first phase project and metadata entry was assigned to experienced information scientists that could safely distinguish among archival and administrative material, identify and remove duplicates (probably scattered across different folders of primary material), perform subject cataloguing, produce abstracts of items (where appropriate) and provide clear guidelines to the digital files processing staff (e.g. for file merging/splitting).

Furthermore, the heterogeneity of the material created a requirement of various ways of browsing to be applied (e.g. not only listings but also tag clouds for presenting subjects and keywords; image-based browsing of material types) and also accessing the digital material itself (e.g. both the ability to download a PDF file and to stream it via an online reader; video material embedded into item pages).

3. Implementation

3.1 Metadata schema

Due to the heterogeneity of the material to be included in the repository, a custom application profile has been developed, utilizing elements from various metadata standards, in particular Dublin Core, Learning Object Metadata (LOM) and PREservation Metadata Implementation Strategy (PREMIS). Furthermore, a new EKT namespace/schema was created to capture custom elements that were required for our implementation but were not available in standard schemata.

The EuroVoc thesaurus [1] has been utilized for the thematic indexing of the material, as well as for spatial coverage. This particular vocabulary was selected due to the following properties:

(a) Breadth and depth of coverage and inter-disciplinarity. EuroVoc is a detailed thesaurus across scientific disciplines and domains, so it is perfectly suitable for the diverse material of our case.

(b) Multi-linguality. EuroVoc is available in 24 European languages. Therefore, the thematic index can be made automatically available in all these languages.

Furthermore, regarding education levels, the controlled vocabulary of the Eurydice network [2] was adopted.

3.2 Content submission and processing workflow

As mentioned above, the content of the repository includes diverse material in terms of their format and type, therefore it was necessary to implement a workflow which would include the required processing of the digital data depending on their type and format.

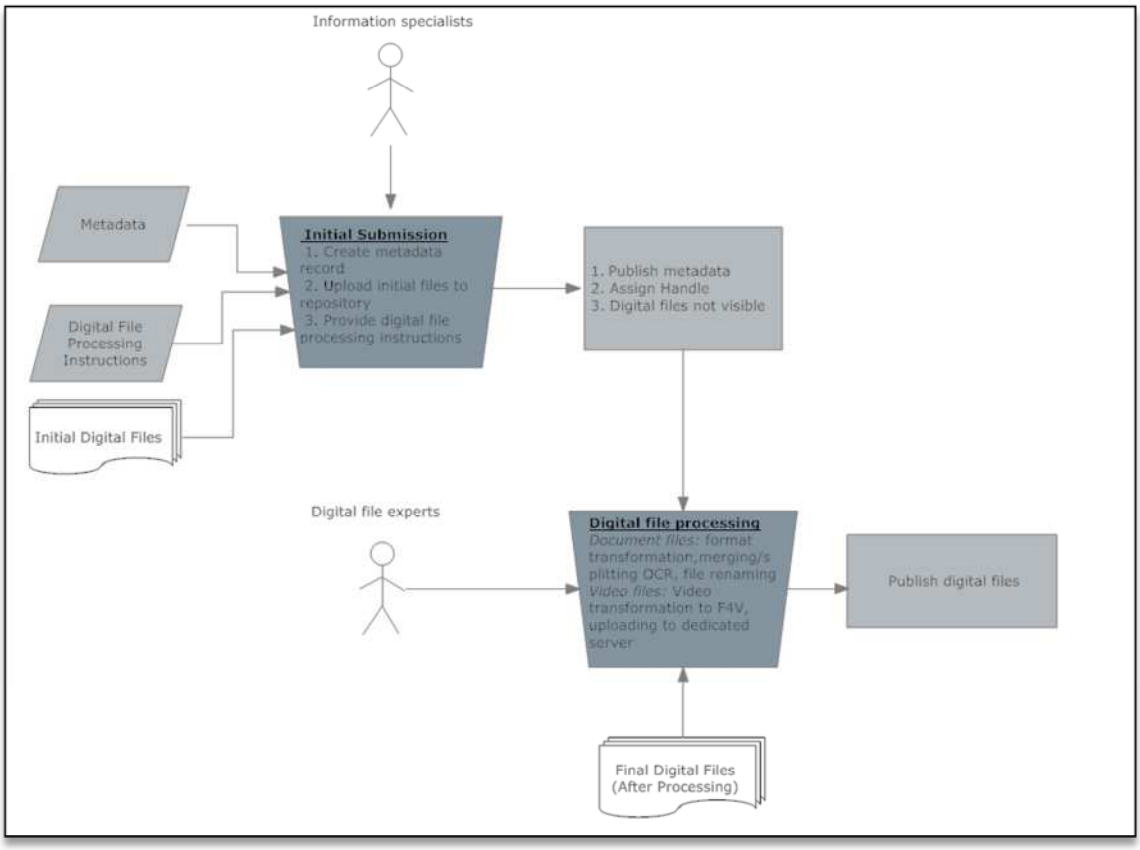


Figure 1 Content submission and processing workflow

The workflow of the repository consists of the following steps, graphically depicted in the aforementioned Figure 1:

1. **Create metadata record:** Firstly, the item is submitted to the repository using an appropriate submit form. This step is performed by authorized users, specialized in the organization and management of information materials (in our case, information specialists who belong to the library staff of EKT/NHRF). In addition to providing the necessary metadata, the information specialists also upload the input digital files and provide, in separate fields and in a custom specification language, the necessary instructions for the processing of the digital file(s) by the digital material processing staff. Specifically:
   a. ***Digital Processing Instructions:*** necessary instructions for the processing of the digital file(s), such as merging sequence for multiple files, renaming instructions, etc.
   b. ***Managing metadata:*** Information regarding the filepath of the specific files (in the material provided from the Special Managing Authority within the Ministry of Education), which can be used as a reference of the cataloguing process status.
2. **Handle assigned to record:** Upon submission by information specialists, the item metadata record is published online and a persistent identifier is assigned to it. At this stage, the digital files are already within the repository, but not visible and available only to authorized personnel.
3. **Digital file processing:** The processing of the digital files from the respective expert staff takes place. It follows the directives of the information specialists for the production of final digital items for publication of each record. In case of text material, the final digital item must comply with the following rules:
   i. The final document must be a full-text indexable and searchable pdf file. Optical character recognition (OCR) might be needed to achieve that, for example when the input file(s) are in pdf format, and they are either image pdfs or have improper encoding or font support (e.g. for Greek content).
   ii. If specified by the information specialist, the final document(s) may be the outcome of merging or splitting the input data files.
   iii. The name of the final document should follow a particular pattern, which includes the item's handle in the repository.

In the case of video material, it has been selected to be hosted by an external FLV streaming video server and integrated to the repository using enhanced FLV players. Video has been received in various formats and the workflow is as follows:

i. The video is encoded in H.264 format and the files are saved in the standard file format.
ii. If specified by the cataloguer and/or considered necessary by the digital material staff, some video editing or enhancement takes place.
iii. The video is uploaded to a the dedicated video streaming server (WOOWZA) and the RTMP link is stored in a separate field in the repository.
iv. The embedded video player is generated dynamically according to the stored RTMP link, which includes also time related information, i.e. start and stop of each video segment corresponding to a serarate metadata entry.

4. **Publication of digital files:** After the processing of digital files the final files become publicly available. Furthermore, the respective staff member records in an appropriate metadata field the fact that the processing has been completed.

4. Advanced user interaction and presentation features

Several advanced user interaction and presentation features, not typically found in repositories, have been implemented to improve the user experience.

The DSpace platform was greatly customized and enhanced in order to accommodate the requirements of this specific digital repository implementation. Furthermore a number of additional software and infrastructure elements, namely video server, image backend server processing and delivery for achieving page by page reading of the digital items were developed. A number of the aforementioned DSpace extensions have been incorporated to the DSpace 3.0 Release. The main issues that have been addressed include modifications in the submission form in order to simplify the process, supporting multi-lingual controlled vocabularies both in metadata entry and presentation to end user and context-sensitive presentation of material as well as a range of modifications in order to encourage efficient and effective human-computer interactions and optimize the overall user experience in the repository.

4.1 Submission form modifications

The default submission process in DSpace platform comprises quite a few steps (collection selection, metadata entry, digital file upload, veriy and grant copyright license). In order to somewhat simplify the process, we changed the submission form so as the user would be able to provide the item metadata and upload one or more digital files in the same step, in a single web page/form. Thus, the submission process is more simple and straightforward, as it consists of three steps: (1) collection selection, (2) describe (metadata entry including copyright, upload one or more files), (3) verify.

Moreover, the following new input-types have been developed: *'advancedName', 'refinementdrop_advancedName', 'refinementdrop_value', 'multiTextArea', 'onebox_lang', 'textarea_lang', 'startHeading'* and *'endHeading'*. These input-types include capabilities like dynamic processing of persons' names which may be entered in various patterns and storing the value as expected by the repository (e.g. *{Surname, First and Middle Names}*), further refine the values by using controlled vocabularies, handle multiple inputs from the same textarea, or allow the user to choose the language of each metadata value.

Finally, in order to improve the user experience and make the submission easier, the submission process was modified in order to be fully localised and support different languages. Specifically, all the labels and control vocabularies in submission form can be configured from the message properties and input-forms configuration file and be displayed in different languages. The item presentation page, also uses these configuration files in order to display the stored control vocabulary in the user's respective language. Also, help tips were implemented which are shown when the user clicks on one of the submission form fields and provide context-sensitive help.

4.2 Visual representation of browsing

The repository browsing functionality was greatly enhanced by applying visual representation mechanisms. The visitor can browse through the subject classification metadata using a tag cloud which provides a quick perception of the most frequently used EuroVoc terms. The tag cloud is fully parameterisable. Specifically, parameterisation concerns how the tag cloud is displayed (e.g. maximum number of display terms, order of appearance rules for terms, minimum required occurences of a term for showing up in the tag cloud, colour and size patterns for tags, etc.) and selecting which metadata fields this browsing should be applied on. Furthermore, a custom, more user-friendly browsing by type has been implemented using a characteristic image per type, in addition to string labels.

The repository homepage, which includes visual representation mechanisms for browsing has been implemented as an extension to DSpace Community. This extension allows the creation of a tag cloud in any repository page for any metadata field (i.e. subject, keyword). The tag cloud is fully configurable via the DSpace configuration files. The tag cloud is implemented as a custom Java Server Pages (JSP) tag, allowing it to be easily included in any page of the DSpace repository besides the homepage. An example fragment of a configuration file is shown below.

```
#########################################################
# TAG CLOUD configuration #
#########################################################
#
# Should display tag cloud in the home page?
# Possible values: true | false
webui.tagcloud.home.show = true
#
# Select the browse index to create a tag cloud for in the home page
# Possible values: any of the browse indices declared earlier in this conf file
webui.tagcloud.home.bindex = subject
#
# Select the total tags to show
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.maxtags = 50
#
# Should display the score next to each tag?
# Possible values: true | false
webui.tagcloud.home.showscore = false
#
# The score that tags with lower than that will not appear in the rag cloud
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.cutlevel = 5
#
# Should display the tag as center aligned in the page or left aligned?
# Possible values: true | false
webui.tagcloud.home.showcenter = true
#
# The font size (in em) for the tag with the lowest score
# Possible values: any decimal
webui.tagcloud.home.fontfrom = 1.3
#
# The font size (in em) for the tag with the highest score
# Possible values: any decimal
webui.tagcloud.home.fontto = 2.8
#
# The case of the tags
# Possible values: Case.LOWER | Case.UPPER | Case.CAPITALIZATION | Case.PRESERVE_CASE |
Case.CASE_SENSITIVE
webui.tagcloud.home.tagcase = Case.PRESERVE_CASE
#
# If the 3 colors of the tag cloud should be independent of score (random=yes) or based on the score
# Possible values: true | false
webui.tagcloud.home.randomcolors = true
#
# The ordering of the tags (based either on the name or the score of the tag)
# Possible values: Tag.NameComparatorAsc | Tag.NameComparatorDesc | Tag.ScoreComparatorAsc |
Tag.ScoreComparatorDesc
webui.tagcloud.home.tagorder = Tag.NameComparatorAsc
#
# The first color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor1 = D96C27
#
# The second color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor2 = 424242
#
# The third color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor3 = 818183
```

4.3 Handling and presenting Controlled Vocabularies

A major feature, needed to accommodate the requirements of this specific digital repository implementation, concerned the support of user-friendly presentation and search for multi-lingual controlled vocabularies. Users need to view and search terms of controlled vocabularies in human-friendly forms and terms need to be available in different languages. At the system level, however, it is sometimes preferable to store encoded values for terms, for the sake of guaranteed unique identification and clarity. For example, in a controlled vocabulary for the “language” metadata field, it is better for the system to store for a language a standard encoded value according to the ISO 693-2 specification (e.g. “eng” for English, “gre” for Greek), while users wish to view the corresponding values as “English” or “Greek” in item metadata pages (or respectively, “Englisch” or “Griechisch” for German users) and get appropriate results both when searching the repository using the queries “English”, “Englisch” or “eng” for the language metadata field.

This functionality has been implemented in the presented repository in which a search for a controlled vocabulary term produces the same set of 1178 results either by searching for items with the English term “Greek” or the respective Greek term “Ελληνικά”, while in the repository database the respective value (“gre”) stored is the code name for the Greek language in the ISO 639-2/B standard.
To achieve this, enhancements to the DSpace repository platform have been necessary. During the submission process, DSpace supported the entry of controlled vocabularies having for each vocabulary term a separate “displayed-value”, a user-friendly one (e.g. “English”), and a respective "stored-value" which is stored in the database (e.g. “eng”). Nevertheless, the only use of a “displayed-value” is in drop-down lists in the submission form; this value would not appear in item metadata pages and would not be indexed for search.
Thus, improvements were applied in order to show the displayed values in the item page and index both stored and displayed values for search. Both extensions are fully configurable via the DSpace configuration files. In particular, the repository administrator can specify if it is desired to show the displayed-value in item page or if DSpace should index both the "stored-value" and the “displayed-value”. Both functionalities have been contributed to the DSpace open source platform and have been incorporated in the core DSpace platform, since version 3.0 [3][4].

4.4 Video integration and external video server system

As already stated, video content forms a considerable part of provided content in the presented repository, whose material includes full length documentaries, short stories for educational use, online courses, etc. Various file formats and encoders were used in the content, from DVD MPEG2 VOD files to proprietary video files and encoding formats (H.263, MPEG-1, MPEG-4, DIVX). On the other hand a unified, meaningfull and intuitive experience should be provided to the end user when accessing this content. Relevant requirements were:

- Easy to use and intuitive user interface for streaming video.
- The option for downloading the video file should not be given, due to IPR issues.
- Integration of a web based video player to the repository and not an external video player program.
- Based on open file standards and codecs, if possible.

In order to cope with the various file formats and encoders the following decision were made:

- Content was batch transcoded to H.264 enconding using a F4V video file container, selected based on the encoder properties and its open nature.
- The video stream was provided by an appropriate video server external to the core repository system. The Woowza video server was selected based on means of versatility in the protocols used (RTMP and RTSP) and efficiency.
- The flowplayer embedded video player was selected based on features and configuration options. The flowplayer configuration is dynamically generated based on the video repository metadata. E.g. in case of a single video stream, with different repository entries, e.g. a DVD with multiple episodes, it was decided that instead of splitting the video file, rich metadata, including episodes start and end time would be included. This approach enables the capability to easily rearrange the video player configuration, be independent of particular video player and streaming servers and provide future enhancements, according to available metadata.

Relevant examples are available at repository.edulll.gr on the video content category, demonstrating the dynamically generated embedded video player according to the video’s repository metadata.

4.5 Providing an ebook-like experience as an alternative to PDF viewer

The initial format of most of the material imported to the repository was PDF files. While PDF is a widespread format, a large percentage of the files included rich photographic material, multihundred pages books, or has been OCR processed, thus file sizes of tens, or even hundreds, of megabytes were common. Therefore an alternative was sought in order to not only make shorter the time required for opening the document, but also to provide a more intuitive user experience.
Initiatives such as the “Google Books” and “Google Art” project, the Internet Archive “Open Library” and advanced repository systems, e.g. Islandora, etc., have paved the way for novel online reading capabilities and experience, with features such as “page by page” reading of electronic resources and tile-based image viewing systems, exploiting advanced codecs such as JP2000.
In order to provide such a viewing experience the following backend components were integrated with DSpace:

- The online reader, based on the open source Internet Archive Open Library BookReader provides the end user interface. It features advanced end user capabilities such as interactive online reading with zooming, thumbnail view, full-text search and hit highlighting.
- A two tier backend for processing, generating and delivering image files, comprising:
  - A distributed multithreaded conversion management tool, in order to interface with DSpace and manage the batch conversion process from PDFs to page by page JP2000 files which wraps around existing standalone converters. This tool can batch convert selected PDF content from a DSpace repository to JP2000, with the conversion process transparently being initiated from the repository manager. It is available as open source software [5] and has been already been integrated also with Open Journal Systems e-publishing platform [6].
  - The highly scalable image delivery backend, based on the The Djatoka image server for providing the page by page content in openurl format and in various sizes. It comprises a Nginx/Varnish load balancer/caching frontend, connected to a 2 node tomcat/djatoka cluster processing on the fly JPEG2000 over a shared SAN storage and utilising a 3 node Postgres 9 cluster for handling file locations and file technical metadata. The overall system architecture is depicted in Figure 2.

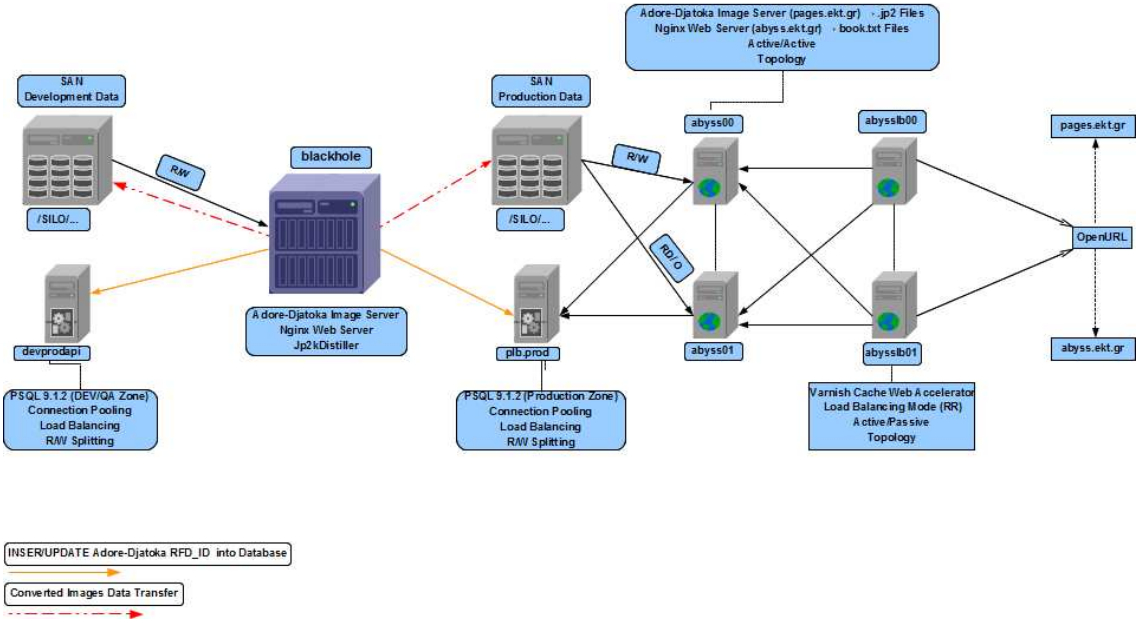


Figure 2 Architecture of the back-end image server infrastructure

The system provides a more natural way to open and view PDF based content, especially suitable for the case of multi-MB PDF files and tablet devices.

An architecture involving a CERIF/CRIS system

The metadata that is available for the digital material (e.g. documents, data sets, multimedia material) is to a considerable extent contextual; thus, it represents the context in which the digital content has been been created, including entities such as persons (e.g. authors, editors), funded projects and the corresponding funding programmes, organizations (e.g. funders, author organisations, project coordinators or supervisors). The context information comprises also the relationships among these entities for example project-person, project-organisation, publication-person, publication-organisation, and many others. Contextual metadata is extremely useful in the case of grey objects, where, compared with white literature, there is no publication venue (e.g. a journal or conference) that is known to the reader and provides some level of reliability guarantees for the content. Therefore, it is even more useful to become aware of provenance and other information providing answers to questions like “is this document produced by a reliable source – e.g. authored or supervised by a reputable person and/or organization?”, “is this document the result of a project in a well-known funding programme?”.

Flat metadata schemata commonly used in digital repositories are not since they do not represent contextual entities as first-class citizens and do not have the capabilities to represent rich semantics on relationships among entities. This kind of metadata is handled very well by Current Research Information Systems (CRIS) and the dominant metadata standard in this area, namely the Common European Research Information Format (CERIF) [7], whose suitability for grey literature has been well documented in other work [8][8].

Therefore, to appropriately represent contextual metadata, a CERIF/CRIS has been developed running as

a separate system than the repository. It is being used to represent in CERIF metadata about persons, projects, funding programmes, organisations and links among those entities and grey objects. The relationship semantics are represented using the CERIF Semantic Layer [7]. Example relationships that need to be represented are project-organisation (indicative roles: funder, supervisor, contractor), person-project (indicative roles: coordinator, team leader, team member, evaluator), person-grey_object (author, numerous types of contributors) and recursive relationships, for example part-of or derived-from relationships among grey objects.
The system is currently being used as a private system that forms an authoritative source of the available contextual metadata, avoiding ambiguity and information loss. Public access is not provided to the CERIF/CRIS, due to considerations originating from the commissioning organization; however, the CERIF/CRIS is very valuable in ensuring the integrity of information that is provided publicly.

Conclusions - future work

In this repository development use case, technical enhancements were used to provide the means to organise, process and import to the repository a wide range of heterogeneous material. Special facilities for the support of these workflows was included, along with enhanced user viewing capabilities for metadata, PDF files and video content, providing an end user experience suitable for the repositoty's users, which include the educational community, researchers and scientists and the general public. Future work, possibly in the frame of a sequel project, includes the inclusion of further material from past funded projects, and the integration of a workflow for the quick inclusion of content produced by currently running projects.

References

[1] EuroVoc thesaurus, http://eurovoc.europa.eu.

[2] The Eurydice Network, http://eacea.ec.europa.eu/education/eurydice/index_en.php.

[3] Show display values for controlled vocabularies in Item Page, DSpace feature request DS-1125 (https://jira.duraspace.org/browse/DS-1225) and pull request #54 (https://github.com/DSpace/DSpace/pull/54).

[4] Index (for search) both display and stored values for fields using controlled vocabularies, DSpace feature request DS-1231 available at https://jira.duraspace.org/browse/DS-1231 and pull request #55 (https://github.com/DSpace/DSpace/pull/55).

[5] dPool Elastic Cluster - PDF/TIFF to JP2000 Distiller, Open source project available on Google Code, http://code.google.com/p/dpool/.

[6] Stathopoulos, P., Houssos, N., Stathopoulou, R., Stavrou, G., & Soumplis, A. (2011). Enhancing OJS journals with advanced online reading and viewing capabilities. Presented at the PKP Scholarly Publishing Conference 2011, http://pkp.sfu.ca/ocs/pkp/index.php/pkp2011/pkp2011/paper/view/319.

[7] Jörg, B. (2010). CERIF: The common European research information format model. Data Science Journal, Volume 9, pp. 24-31.

[8] Jeffery, KG, Asserson, A. (2009). Mosaic: Shades of Grey. Conference Proceedings on Grey Literature: GL-conference series, ISSN 1386-2316 ; No. 11.

[9] Jeffery, KG, Asserson, A. (2011). GL Transparency: Through a Glass Clearly. The Grey Journal (TGJ), ISSN 1574-1796, Volume 7, Number 2.

https://doi.org/10.26069/greynet-2026-000.548-tgj

Customized OAI-ORE and OAI-PMH Exports of Compound Objects for the Fedora Repository *

Alessia Bardi, Sandro La Bruzzo, and Paolo Manghi (Italy)

Abstract

Modern Digital Library Systems (DLSs) are based on document models which surpass the traditional payload-metadata document model to incorporate further entities involved in the research life-cycle. Such DLSs manage graphs of interconnected objects, hence offer tools for the creation, visualization and exports of such graphs. In particular, objects in the graph are exported via standard OAI-ORE and OAI-PMH protocols, encoded as (XML) "packages of interlinked information objects", also known as compound objects. Fedora is a well-known repository platform, designed to support the realization of DLSs implementing modern document models. To date, Fedora does not provide tools to customize compound object exports from DLS object graphs. This paper presents Fedora-OAIzer, an extension of Fedora which allows DLS developers to customize the structure of compound objects to be exported from a given DLS document model – expressed in terms of Fedora Content Models – and to select the OAI protocol of preference. In order to prove the completeness of the approach, Fedora-OAIzer is compared to other solutions for exporting compound objects from Fedora repositories.

1 Introduction

In the past, Digital Library Systems (DLSs) adopted "traditional" document models representing collections of payloads of digitized or born-digital material (e.g., publications, multimedia files) and their digital descriptions (i.e., metadata records). "Modern" document models enhance the traditional document model to incorporate further entities involved in the research life-cycle and semantic relationships. Consequently, modern DLSs manage graphs of interconnected information objects, called *object graph*, rather than flat collections of objects. For example, a "traditional" publication-metadata document model can be enriched with further contextual information, such as the used and generated research data.

The complexity of modern document models introduces a number of challenges concerning the way graphs of information objects are displayed, encoded, and exported across different DLSs. In particular, sub-parts of the object graph are exported, rather than each single information object or the object graph in its whole, in order to disseminate a package containing semantically related information objects. Such packages, called *compound objects*, are usually encoded in XML or other machine-readable formats and exported via standard protocols. The most used standard protocols are promoted by the Open Archives Initiative [6]: the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [4] and OAI-ORE (Open Archives Initiative Object Reuse and Exchange) [5].

The Fedora Repository [3] is a well-known platform supporting the realization of DLS. Its data model is designed to represent modern DLS document models thanks to its graph-oriented primitives. At the time of writing, Fedora features a non-customizable OAI-PMH publisher[1] which exports only Dublin Core metadata records. An additional module supports the OAI-PMH exports of datastreams with different XML formats. Two plugins for OAI-ORE export are also available online: oreprovider[2] and Fedora2ORE[3] . Both plugins do not allow to fully customize the structure of the exported compound object.

This paper presents Fedora-OAIzer, an extension of Fedora for the export of compound objects conforming to a given portion of the underlying DLS document model through the OAI-PMH or OAI-ORE protocols. Fedora-OAIzer implements a mechanism based on the concept of "OAI view" of a Fedora document model. An OAI view is the sub-structure of the document model that developers select to customize the shape of the compound objects to export. OAIzer interprets OAI views to automatically deploy web APIs capable of exporting compound objects compatible with the given structure and according to the preferred OAI protocol.

Outline

Section 2 introduces basic concepts and defines the addressed problem. Section 3 describes the Fedora repository platform and the available tools for exporting compound objects via OAI-PMH and OAI-ORE. Section 4 presents Fedora-OAIzer and its approach that enables developers to customize compound

* First published in the GL14 Conference Proceedings, February 2013.

objects by selecting sub-parts of the document model at hand. We conclude and compare Fedora-OAIzer to other existing solutions in Section 5.

2 Document models, Digital Library Systems and compound objects

A Digital Library System (DLS) [2] is a DL-oriented software serving a particular DL community. A DLS offers functionalities for the management, access and dissemination of graphs of information objects whose structure is defined by a data model called document model. A document model is a formal definition of types of entities and relationships that a Digital Library (DL) wants to manage. An entity type typically describes properties of objects in terms of name, cardinality and value type. A relationship type usually include a semantic label expressing the nature of the association and the types of entities allowed as sources and targets of the relationship. Figure 1 shows a document model defining three types of entities (Article, PDF, and Data) and the available relationships (HAS PDF, USES, USED BY, GENERATES, GENERATED BY).

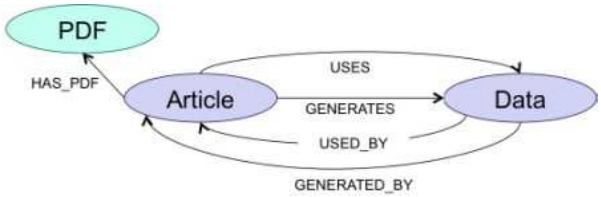


Fig. 1. A modern document model: articles linked with research data

For example, a DLS adopting the document model in fig.1, manages graphs of information objects as that in Fig. 2.

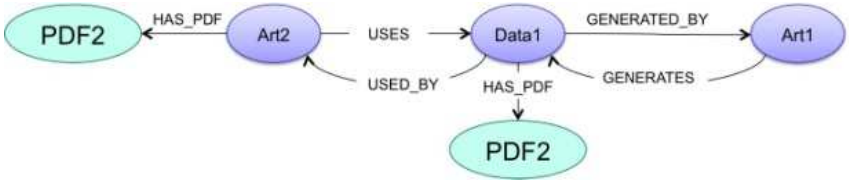


Fig. 2. An example of object graph

To clarify, it is possible to compare the above concepts with similar concepts in the relational database world. An entity-relationship model in the database world corresponds to a DLS document model. DLSs are comparable to applications realized on top of Relational Data Base Management Systems.

Since DLSs manage graphs of information objects, it is important to define the granularity of the data to export. Indeed, exporting each single information object separately (for example, exporting Art1 of Fig.2 without the generated data) leads to a loss of contextual information. Related information objects should be packaged and exported together as a single compound object, capable of maintaining all semantic relationships involving the packaged objects.

In order to exchange and re-use compound objects, interoperability issues must be tackled. The Open Archives Initiative [6] defines two standard protocols for data interoperability and information reuse: OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting [4] and OAI-ORE Open Archives Initiative Object Reuse and Exchange [5]. Both protocols allow to export compound objects of a DLS, but they adopt two radically different approaches. OAI-PMH is meant for the export of the descriptive metadata in the form of XML files; OAI-ORE is meant for the export of web-interpretable RDF representations of so-called aggregations, which are special web resources encoding compound objects.

OAI-PMH provides an application-independent interoperability framework based on metadata harvesting. Its data model has four main elements:

- Resource: an object described by one or more metadata records.
- Metadata record: XML data describing a resource. Each metadata record has its metadata format, often referred to as the XML Schema.
- Item: container of metadata records describing one resource. Each item must have at least one Dublin Core metadata record.
- Set: optional element used to group items.

In order to export via OAI-PMH, DLSs set up an OAI-PMH publisher capable of exporting a set of XML files encoding compound objects. Well-known metadata formats that can be adopted are METS[9] and XML-DIDL[8].

OAI-ORE defines standards for the description and exchange of Web resources called aggregations. An aggregation is a Web resource with its own identity and it represents a group of related Web resources. The OAI-ORE protocol does not define an exchange protocol, but only a data model, while suggesting exchange formats such as XML/RDF and ATOM feeds. The OAI-ORE model captures four type of resources:

- Aggregation: resource that groups other resources, called aggregated resources.
- Aggregated resource: resource that belongs to an aggregation, that is the ORE representation of an information object in a compound object.
- Resource map: serializable description of an aggregation. A resource map:
  - lists the aggregated resources;
  - has properties about the aggregation and its aggregated resources, e.g., relationships among aggregated and other external resources.
- Proxy: resource that allows to assert relationships among aggregated resources in the context of one specific aggregation.

Among the functionalities offered by a DLS, we usually find support for data exchange according to the OAI standard protocols. However most DLSs do not provide means to customize the structure of the compound objects, but they rather fix statically the mapping between the document model and the OAI-PMH and OAI-ORE data models.

3 Fedora and exports of compound objects

Fedora (Flexible Extensible Digital Object Repository Architecture) is a well-known repository platform, designed to support the realization of DLSs. Its data model supports the representation of labelled graphs of information objects. Fedora manages objects of different kinds. “Fedora Data Objects” (FDOs) represent information objects. A FDO is composed by the following parts: an XML Dublin Core metadata record, a list of local or remote files called “datastreams”, and a list of relationships to other FDOs. The latter are serialized as RDF/XML[7] into a special datastream called RELS-EXT. “Fedora Content Models” (FCMs) are special objects devised for the definition of a document model in a Fedora instance. FCMs define constraints on the structure of FDOs, declaring which are the mandatory datastreams, relationships and operations (i.e., Web Service methods) of the FDOs compliant to that FCM. Figure 3 shows how the document model in Fig. 1 can be represented in terms of FCMs. CM article is the content model for the class article and defines one mandatory datastream called ART of mime type PDF. CM data is the content model for the class data and defines two mandatory datastreams: one for binary content called DATA, the other for the XML description of the data in DDI format[1]. By default, Fedora also includes one mandatory XML datastream called DC for metadata in Dublin Core format. The arrows between the two models represent the available semantic relationships as they are declared in the ONTOLOGY datastream of both content models. Listing 1.1 is an excerpt of the ONTOLOGY datastream for CM article.

Listing 1.1. Excerpt from the FCM for the Article class: the ONTOLOGY datastream

```
<!-- ONTOLOGY datastream for allowed relationships -->
<foxml:datastream CONTROL_GROUP="X" ID="ONTOLOGY" STATE="A" VERSIONABLE="true">
  . . .
      <rdf:RDF>
        <owl:Class rdf:about="ns:CM_article">
          <!-- Objects of this class can have the following relations -->
           <owl:ObjectProperty rdf:about="uses"/>
          <owl:ObjectProperty rdf:about="generates"/>
          <!--Both relations must have objects compliant to the CM_data content model
              as targets. -->
          <rdfs:subClassOf><owl:Restriction>
            <owl:onProperty rdf:resource="#uses"/>
             <owl:allValuesFrom
               rdf:resource="ns:CM_data"/>
            </owl:Restriction></rdfs:subClassOf>
          <rdfs:subClassOf><owl:Restriction>
          <owl:onProperty rdf:resource="#generates"/>
               <owl:allValuesFrom
                 rdf:resource="ns:CM_data"/>
           </owl:Restriction></rdfs:subClassOf>
        </owl:Class>
      </rdf:RDF>
. . .
```

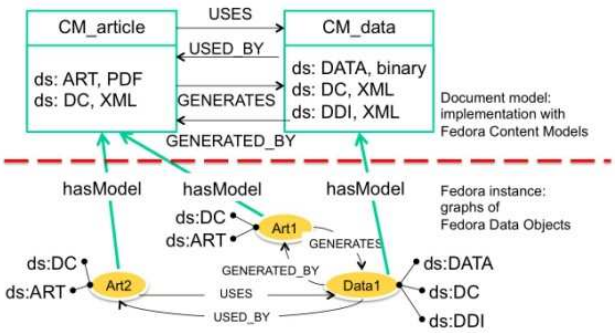


Fig. 3. Fedora content models and data objects example

3.1 Fedora and OAI

The Fedora data model is an expressive data model because its primitives allow to represent graph of information objects without imposing pre-defined structural or semantic constraints. Nevertheless, boundaries of compound objects are fixed, because Fedora's concept of compound object is that of an aggregation of datastreams. This means that in Fedora the notion of compound object matches the definition of a Fedora Object. Given the instance in Fig. 3, each of Art1, Art2 and Data1 are considered by the system as distinct compound objects. Considering a set of interconnected Fedora Objects as one compound object is not possible in Fedora without the realization of a new logic layer on top of it.
The rest of this section describes four existing solutions for the export of compound objects from a Fedora instance.

OAI-PMH Providers

Basic OAI-PMH Provider[4] is the built-in OAI-PMH provider for Fedora. It exports the mandatory DC datastream of each FDO. For the export of datastreams with other metadata formats, then the additional OAI Provider Service[5] module is required. OAI Provider Service supports any metadata format available through datastreams and interprets relationships with a given name to set up OAI Sets. Both tools provide a static mapping from the Fedora data model to the PMH data model. A FDO is mapped into an OAI-item, XML datastreams are mapped into metadata records.

OAI-ORE Providers

OREprovider[6] enables the export of FDOs as OAI-ORE aggregations with an object-oriented approach. It implements a static mapping from the Fedora data model and the OAI-ORE data model. Datastreams are mapped into ORE aggregated resource. For the generation of ORE aggregations there are two modes available. In "annotation mode" FDOs must be annotated with relationships that assert the identifier of the target ORE aggregation and the datastreams to be mapped into aggregated resources. In "auto-creation" mode each FDO is mapped into one ORE Aggregation, whilst each of its datastreams is mapped into one aggregated resource. In "autocreation mode" the tool is easy to use and no alteration has to be done to an existing repository. However, if a DLS developer wants to have control over the exported objects, FDOs must be appositely annotated, hence the document model must be aware of the special relationships exploited by OREprovider. Furthermore, the static mapping does not include the concept of ORE proxy and the tool does not provide any support for the modelling of relationships among aggregated resources.

The Fedora2ORE[7] tool adopts a navigation-oriented approach for the export of FDOs as OAI-ORE aggregations. ORE aggregations consist of one FDO together with the objects reachable by navigating its relationships up to a given depth. Fedora2ORE traverses the object graph starting from an object with a given identifier according to a variant of the breadth first search. A resource map is created to represent the visited sub-graph. The behaviour of the traversal can be customized by statically specifying in a configuration file which relationships, datastreams, FDOs are to be ignored. Each node of the resulting sub graph is an Aggregated Resource. Fedora2ORE is independent from DLS applications because there is no need to define special relationships as for OREprovider. It generates aggregations by following relationships between objects, but DLS developers can only define the boundaries of aggregations in terms of navigation depth rather than in terms of their preferred document model sub-structure. Furthermore, relationships among FDOs are not mapped into the OAI-ORE model, hence most of the semantics of the compound object is lost.

4 Fedora-OAIzer

Fedora-OAIzer is an extension for Fedora that allows to customize exports of compound objects via OAI-ORE and OAI-PMH protocols. As shown in Fig. 4, Fedora-OAIzer realizes a new layer on top of Fedora, capable of dynamically deploying OAI-PMH or ORE-ORE interfaces.

Fedora-OAIzer exploits the information about the Fedora Content Models to construct a representation of the document model of the DLS. We denote such a representation as *entity graph*. DLS developers can choose granularity, shape and properties of the compound objects to export by selecting interesting nodes and edges from the entity graph. Since the entity graph is a representation of a Fedora document model, the selected parts are a sub-structure of the document model. That sub-structure is called *OAI view* of a Fedora document model. The OAI view corresponds to the structure of the compound objects to export. OAIzer interprets OAI views to automatically setup OAI-ORE and OAI-PMH repositories. Repositories are here intended as ORE or PMH APIs available at a given URL, dynamically generated after an OAI view interpretation.

OAIzer exploits the built-in capabilities of Fedora, namely the REST APIs and the triple store, hence it is possible to plug Fedora-OAIzer in any standard Fedora instance.

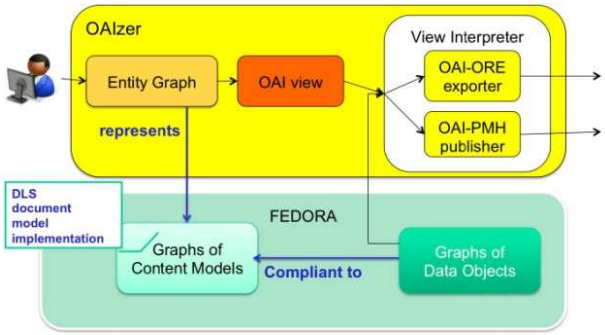


Fig. 4. View mechanism and OAI tools

4.1 Generation of the entity graph

When a DLS is based on Fedora, then developers define the document model in terms of Fedora Content Models (FCMs). Figure 5 shows an example of entity graph representing the document model of Fig.3: nodes represent classes of information objects, that is Fedora Content Models. Nodes are annotated with information about datastreams. Edges represent allowed relationships between objects of the connected classes.

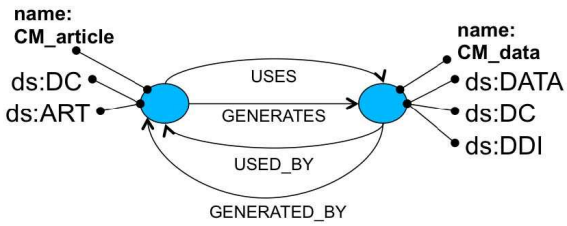


Fig. 5. An example of entity graph

Fedora-OAIzer generates the entity graph by performing the following macro steps:

1. Creation of one node of the entity graph for each FCM in the Fedora instance. The list of existing FCMs can be obtained by querying the triple store[8].
2. Obtain the full XML representation of each FCM using the REST API[9]. Information in that file is used in the next steps.
3. Enrichment of nodes with properties. Properties of a node reflect the declarations of datastreams in the corresponding content model. Such information is extracted from the standard XML datastream named DS-COMPOSITEMODEL.
4. Generation of edges. If the FCM declares a relationship to another content model, then an edge is created between the nodes corresponding to the content models involved in the relation. The edge is labelled with the name of the relationship as it is declared in the ONTOLOGY datastream (see Listing 1.1).

4.2 Definition of the view

OAIzer provides a graphical user interface where DLS developers can see the generated entity graph and define their OAI view, that is the shape of the compound objects to export by choosing the interesting nodes, properties, and edges.

The DLS developer first chooses the root node of the OAI view. The view interpreter navigates the Fedora object graph starting from every object compliant to the content model represented by the root node. After the selection of the root, the DLS developer performs iteratively the following steps until the OAI view is completed according to requirements:

- select one or more properties of the current node. The property names match the names of the corresponding Fedora datastreams: by selecting a property, the developer includes the corresponding datastream in the ORE resource relative to the current node.
- select one or more edges from the forward star of the current node. Each edge is labelled with the name of the corresponding Fedora relationship. If the DLS developer selects an edge, the target node is also included in the OAI view and an ORE relation is added between the ORE resources corresponding to the current and the target node.

The OAI view is eventually serialized into a formal language for the view interpreter.

Figure 6 shows a possible OAI view defined over the entity graph in Fig. 5. Compound objects are rooted in the class CM article and include the DC datastream of the article, the DC and DDI datastreams of the research data objects reachable through relationships labelled with USES, and the DC datastream of articles reachable from those research data objects via relationships labelled GENERATED BY.

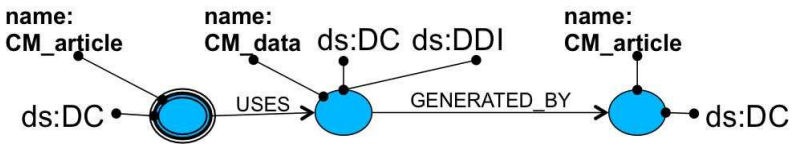


Fig. 6. An example of OAI view

4.3 Interpretation of the view

The View Interpreter is the component of OAIzer that deploys OAI-PMH and OAI-ORE APIs for the dissemination of compound objects whose structure is defined by an OAI view. The OAI view defines a "root" content model: each object compliant to the root content model is the entry point for an *instance of the OAI view.* An instance of the view is a sub-graph of the object graph. It includes all objects and relationships of one compound object.

The interpreter performs a visit on the FDO's graph for each entry point in order to get all FDOs that form an instance of the view. Moreover, the interpreter processes the FDOs of the view instance to keep track of the semantic relationships between them.

Figure 7 shows two instances of the view in Fig. 6 over the Fedora instance in Fig. 3.

For the customization of an OAI-PMH publisher, the interpreter maps the view into an OAI-PMH Set. Each instance of the view is mapped into an OAI-PMH Item with at least two metadata formats: OAIzer-XML and DC. OAIzer-XML is an idiosyncratic format for the representation of compound objects. OAIzer provides a default mapping from OAIzer-XML to DC in order to generate the DC records and be fully compliant to the OAI-PMH protocol. More metadata formats can be added by providing customized XSLT transformations from OAIzer-XML to the target metadata formats (see fig.8).

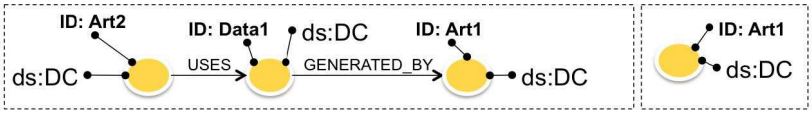


Fig. 7. Instances of the OAI view

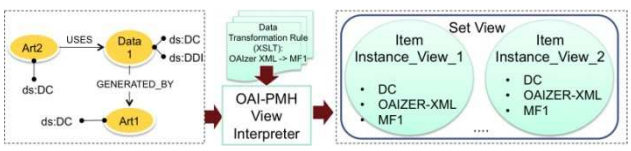


Fig. 8. Examples of OAI-PMH exports

For the customization of the OAI-ORE exporter, the interpreter creates for each instance of the view one ORE aggregation together with its corresponding resource map. For each FDO of the instance, an aggregated resource and an associated proxy are created. The aggregated resource is the OAIzer-XML

representation of the FDO, including the datastreams selected in the view. Finally, the interpreter processes the FDOs of the view instance to add semantic relationships between aggregated resources. At this aim, the interpreter exploits the ORE proxy entities. The interpreter connects two proxies with a relationship r if the Fedora triple store contains the triple: fdo1 r fdo2 where fdo1 and fdo2 are the identifiers of the FDOs corresponding to the ORE proxies (see Fig. 9)

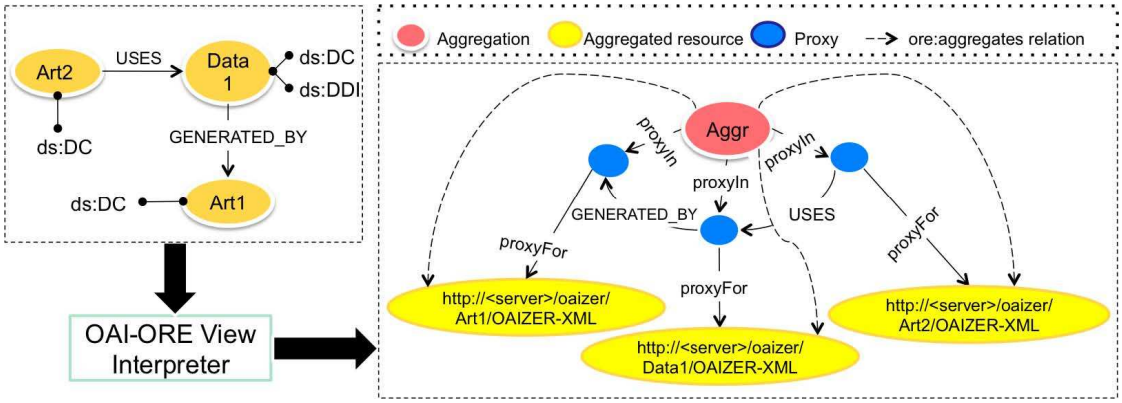


Fig. 9. Examples of OAI-ORE Aggregations

5 Conclusion

We discussed how the evolution of today's Digital Library Systems (DLSs) and the complexity of document models to represent led to data interoperability challenges. We addressed issues regarding the exchange of compound objects, i.e., packages of interlinked information objects with their own identity, among different DLSs via the standard protocols OAI-ORE and OAI-PMH.

We presented Fedora-OAIzer, a tool for the customization of OAI-PMH and OAI-ORE exports of compound object from Fedora repositories. Fedora-OAIzer creates a layer on top of Fedora in order to allow DLS developers to select the shape and boundaries of the compound objects to export. Fedora-OAIzer analyses a Fedora instance and infers the document model at hand from the Fedora content models. The document model is represented as an entity graph from which developers can select a subset of nodes and edges to define the OAI view. The OAI view defines the structure of the compound objects to export. Fedora-OAIzer then interprets the OAI view and, on demand, deploys OAI-ORE and OAI-PMH APIs delivering compound objects whose structure complies with the view definition.

Figure 10 summarizes the comparison among Fedora-OAIzer and the other existing solutions for OAI-PMH and OAI-ORE exports we described in Sect. 3.1.

	OAIzer	Basic OAI-PMH Provider	OAI Provider	OREProvider	Fedora2ORE
Fedora Built-in	NO	YES	NO	NO	NO
Supported Protocols	PMH - ORE	PMH	PMH	ORE	ORE
PMH Item	OAI View Instance	FDO	FDO	n.a	n.a
PMH-Metadata Records	Generated from OAI View Instance	Datastream	Datastream	n.a	n.a
PMH Metadata Format	DC, OAIzer-XML	DC	any format existing datastream	n.a	n.a
ORE Aggregation	OAI view instance	n.a	n.a	definded by annotation	subgraph visited starting from a given FDO
ORE Aggregated Resources	FDOs in the OAI view instance	n.a	n.a	datastreams with a given name	FDOs in the visited subgraph
ORE Proxy	FDOs in the OAI view instance	n.a	n.a	not supported	not supported
Relationships between Aggregated Resources	YES	n.a	n.a	NO	NO
Compound object boundaries	OAI view (properties and navigational criteria)	FDO	FDO	FDOs annotated with the same tag	navigation depth

Fig. 10. Comparing Fedora-OAIzer to other OAI solutions for Fedora

Acknowledgements
This work is partly funded by the European Commission as part of the projects OpenAIRE (FP7-INFRASTRUCTURES-2009-1, Grant Agreement no. 246686) and OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595).

References

1. D. D. I. Alliance. Data Documentation Initiative. http://www.ddialliance.org/.
2. L. Candela, D. Castelli, P. Pagano, C. Thanos, Y. E. Ioannidis, G. Koutrika, S. Ross, H.-J. Schek, and H. Schuldt. Setting the foundations of digital libraries: The delos manifesto. D-Lib Magazine, 13(3/4), 2007.
3. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An Architecture for Complex Objects and their Relationships. Journal of Digital Libraries, Special Issue on Complex Objects, 2005.
4. C. Lagoze and H. Van de Sompel. The OAI Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html.
5. C. Lagoze and H. Van de Sompel. The OAI Protocol for Object Reuse and Exchange. http://www.openarchives.org/ore/.
6. C. Lagoze and H. Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, pages 54–62. ACM Press, 2001.
7. F. Manola and E. Miller. RDF primer. Technical report, W3C Recommendation, 2004. http://www.w3.org/TR/rdf-primer/.
8. MPEG-21, Information Technology, Multimedia Framework. Part 2: Digital Item Declaration. Technical report, ISO/IEC 21000-2:2003, 2003.
9. The Library of Congress. Metadata Encoding and Transmission Standard. http: //www.loc.gov/standards/mets/, February 2002.

[1] Basic OAI-PMH Provider, https://wiki.duraspace.org/display/FEDORA36/Basic+OAI-PMH+Provider
[2] oreprovider Fedora module, http://oreprovider.sourceforge.net /
[3] Fedora2ORE, http://trac.eco4r.org/trac/eco4r/wiki/Fedora2ORE
[4] https://wiki.duraspace.org/display/FEDORA35/Basic+OAI-PMH+Provider
[5] https://wiki.duraspace.org/display/FCSVCS/OAI+Provider+Service+1.2.2
[6] http://oreprovider.sourceforge.net/index.html
[7] http://trac.eco4r.org/trac/eco4r/wiki/Fedora2ORE
[8] select ?o from <#ri> where
?o <fedora-model:hasModel>
<info:fedora/fedora-system:ContentModel-3.0>
[9] http://<fedoraServer>/objects/<id>/objectXML

On the News Front

Report to the ARC Grey Literature Project on the Fourteenth International Conference on Grey Literature

Gerald White (Australia)

I was privileged to be able to attend the 14th International Conference on Grey Literature (GL14) titled 'Tracking Innovation through Grey Literature' in Rome on 29th and 30th November, 2012. The purpose of attending the conference was to deliver a presentation and a paper as well as to network internationally as a part of the Australian Research Council (ARC) research project Grey literature, innovation and access to knowledge: realizing the value of informal publishing.

GL14 conference

The GL14 conference (http://www.textrelease.com/gl14conference.html) was held at the Italian National Research Council's (CNR) Rome headquarters and attracted 100 researchers world-wide, although notably, most were from Europe. My wife and I were indeed fortunate to be invited to attend with the conference organisers for the grey literature award dinner at the magnificent Forum Hotel, on the evening prior to the conference. Dr. Petra Pejšová, manager of the Czech Republic national repository of grey literature was awarded the GreyNet Award 2012 for her leadership of an outstanding national service.

The definition of grey literature, determined some years earlier, was not debated by the speakers at the conference but accepted as the conventional construct. Grey literature is 'information produced on all levels of government, academia, business and industry in electronic and print formats but which is not controlled by commercial publishers and where publishing is the not the primary activity of the producing body'. However, the use of different formats and applications of digital technologies were at the forefront of considerations and discussions because much literature and information today is born digital and remains digital for its useful lifetime.

The two main issues which emerged from the conference and that have yet to be resolved with grey literature were the development of agreed processes for determining information quality, and its management and preservation. That is, in a world where information and literature are born digital and remain digital, how do authorities decide what information to archive and what information should be made readily accessible or preserved for future research?

The alternative to the management and preservation of grey literature is the wastage or loss of knowledge in the form of research, expert commentary and reports that have cost large sums of money to develop. There are many examples in government, academia, business and industry of where this has happened over the last two decades. My paper and presentation demonstrated how a decade of research, reporting and expert commentary of a major national education and training service costing millions of dollars had been negligently but deliberately lost to posterity. An appreciation in Australia by government, academia, business and industry of the importance of grey literature could alleviate this dilemma.

GreyNet

GreyNet (www.greynet.org) is an international membership based grey literature network service that includes a journal and a range of useful databases. GreyNet manages and hosts the grey literature annual conferences (www.textrelease.com) with the assistance of an international advisory group of grey literature experts. The network of experts involved in the grey literature network provides a rich source of experience and expertise in the field. Also available from GreyNet are a regular newsletter and advice compiled in the book Grey Literature in Library and Information Studies by D. Farace and J. Schopfel (Eds) (2010) www.textrelease.com/publications/monograph.html as well as via GreyNet's helpdesk and Referral service www.greynet.org/home/helpdesk.html.
The GreyNet International has now been in operation for over fifteen years and has extensive experience and knowledge in the area of grey literature.

Exemplary services
Regions that have successfully taken steps to develop and implement systems for the management and preservation of grey literature include the European Union, France, Italy and the Czech Republic. The services in each of these regions are briefly mentioned below.

European Union
A European repository of openly accessible research articles called OpenAIRE (www.openaire.eu) is available for the deposit, storage and retrieval of research articles in the open domain. OpenAIRE is the storage and management vehicle for the outputs of research grants for 32 European countries. Final copies and published copies of research grant documents are required to be deposited onto this system by grant recipients.

France
Grey literature, on the other hand, can be stored in the OpenGrey Repository hosted by Institut de l'Information Scientifique et Technique, INIST/CNRS (www.opengrey.eu) which contains close to 1 million documents and includes research reports, doctoral dissertations, conference papers and more in the sciences and humanities. OpenGrey has integrated the database records of the former System for Information on Grey Literature in Europe (SIGLE) together with the conference proceedings of the grey literature conferences, GreyNet as well as OpenSIGLE. This system has a 25 year history and is well established in Europe.

Italy
The GL14 conference was held at the premises of the publicly funded Italian National Research Council (CNR) which focusses on technical scientific and social population research. CNR houses an extensive database of information that includes published and grey literature available for its 8,000 staff that includes 4,000 researchers.

Czech Republic
The Czech Republic has developed a robust digital library of grey literature called the National Technical Library (www.nusl.cz) or NTK. NTK is a very sophisticated and robust grey literature service that enables contributions of grey literature from information producers. With over 200,000 records in science, research and education, NTK has extensive global partnerships and a comprehensive set of policies for the systematic collection of metadata and documents. Support for grey literature occurs through annual NTK workshops (nrgl.techlib.cz/index.php/Workshop), excellent public information about the NTK service (nrgl.techlib.cz/index.php/About_NRGL) and a useful publication called Grey Literature Repositories by Marcus Vaska et al. (2010) (http://nrgl.techlib.cz/index.php/Book).

Conference Proceedings
The conference proceedings containing the full set of papers are to be released in February 2013 and will be available from the grey literature conference site (www.textrelease.com). It can be purchased in print, CD or online formats for 99 Euros. A copy of the final paper that I presented will be included although it is also attached to this report.

GL15 Conference 2013
The next conference in 2013, the Fifteenth International Conference on Grey Literature has been titled The Grey Audit: A Field Assessment of Grey Literature and will be held on 2nd and 3rd of December in Bratislava at the Slovak Republic's Centre of Scientific and Technical Information. GL15 will mark the 20th Anniversary of the International Conference Series on Grey Literature 1993-2013.

Summary and recommendation
A number of noteworthy international projects have been developed over the last decade including GreyNet International, the work of the European Union with OpenAIRE, the OpenGrey Repository, as well as the NTK grey literature repository of the Czech Republic. NTK includes extensive guidelines about grey literature, comprehensive details about metadata, and information about document selection and preservation.

I would recommend that during the life of the ARC grey literature research project a person from the project or its partners be supported to attend the next international GL-Conference. Attendance at the conference will help to maintain and build an international network that will assist in keeping our research project abreast of global projects and trends in grey literature.

https://doi.org/10.26069/greynet-2026-000.548-tgj

Fifteenth International Conference on Grey Literature

The Grey Audit: A Field Assessment in Grey Literature

Slovak Centre of Scientific and Technical Information

CVTI SR, Bratislava, Slovak Republic, 2-3 December 2013

Call for Papers

Guidelines for Abstracts

Title of Paper:	Topics/Themes:
Author Name(s):	Telephone:
Organization(s):	Fax:
Address/P.O. Box:	Email:
Postal Code - City - Country:	URL:

Participants who seek to present a paper at GL15 are invited to submit an English abstract between 350-400 words. The abstract should deal with the problem/goal, the research method/procedure, an indication of costs related to the project, as well as the anticipated results of the research. The abstract should likewise include the title of the proposed paper, topic(s) most suited to the paper, name(s) of the author(s), and full address information. Abstracts are the only tangible source that allows the Program Committee to guarantee the content and balance in the conference program. Every effort should be made to reflect the content of your work in the abstract submitted. Abstracts not in compliance with the guidelines will be returned to the author for revision.

GL15 Conference Topics and Related Themes:	
Technology Assessment	*Digital Preservation, Design and Populating Repositories, Open Source*
Investing in Standards	*Metadata, Resource Sharing, Peer Review, and Curation*
Research and Data	*CRIS, Data Management, Research Communities, Scientific Discovery*
Towards Informed Policies	*Economic, Legal, and Ethical Concerns, Governance, Open Access*
Sustaining Good Practices	*Guidelines, Case Studies, Education and Training*
Trends and Breakthroughs	*Crowd Sourcing, Creative Workflows and Workplaces, Engines of Change*

Due Date and Format for Submission

Abstracts in MS Word must be emailed to conference@textrelease.com by April 8th 2013. The author will receive verification upon receipt. Shortly after the Program Committee meets on April 26th 2013, authors will be notified of their place on the conference program. This notice will be accompanied by further guidelines for submission of full text papers, PowerPoint slides, as well as Conference Registration.

GL15 Program and Conference Bureau
TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com ▪ conference@textrelease.com
Tel/Fax +31-20-331.2420

https://doi.org/10.26069/greynet-2026-000.548-tgj

Fifteenth International Conference on Grey Literature

The Grey Audit: A Field Assessment in Grey Literature

Slovak Centre of Scientific and Technical Information

CVTI SR, Bratislava, Slovak Republic, 2-3 December 2013

Registration Form

The Conference Registration Fee includes attendance during the Sessions, a copy of the GL15 Program and Abstract Book, the full-text of the Conference Papers and PowerPoint Slides, as well as the conference pouch and participant badge. Also included are lunches, refreshments during the morning and afternoon breaks, and the Inaugural Reception.

Payment after October 1st 2013 add a €75 surcharge

☐ **€ 495** Participant Fee ☐ **€ 395** Author Fee ☐ **€ 250** Day Pass ☐ **€ 100** Student Fee

I am employed with **CVTISR** and entitled to a 20% reduction on the conference fee: ☐

I am a *GreyNet* Member and am entitled to a 20% reduction on the conference fee: ☐

Registrant's Name: _____

Organization: _____

Address/P.O. Box: _____

Postal Code/City/Country: _____

Tel/Fax/Email:

Please Check One of the Boxes below for the Method of Payment:

☐ Direct bank transfer to TextRelease, Account 3135.85.342 - Rabobank, Amsterdam, Netherlands
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to GL15 and Registrant's Name

☐ MasterCard ☐ Visa Card ☐ American Express

Card No. _____ Expiration Date: _____

If the name on the credit card is not that of the registrant, print the name that appears on the card, below

_____

Signature _____ **CVC II code** _____ (*Last 3 digits on signature side of card)*

Place _____ Date _____

Note: Credit Card transactions can be authorized by Phone, Fax, or Postal services. Email is not authorized.

GL15 Program and Conference Bureau
TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com · **conference@textrelease.com**
Tel/Fax +31-20-331.2420

Assante, Massimiliano
Massimiliano Assante holds a master degree (M.Sc.) on Information Technologies received from the University of Pisa, where he is undertaking a Ph.D. on Information Engineering. He is research staff at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" (ISTI), an institute of the Italian National Research Council (CNR). He joined ISTI in 2007 and is currently member of the iMarine EU Project and EUBrazilOpenBio Project. In the past he has been member of D4Science II, D4Science, DILIGENT and DRIVER European Projects. His research interests include Data Infrastructures, Next Generation Digital Libraries, Information Systems and NoSQL Data Stores. Email: massimiliano.assante@isti.cnr.it

Bardi, Alessia
Alessia Bardi received her MSc in Information Technologies in the year 2009 at the University of Pisa, Italy. She is a PhD student in Information Engineering at the Engineering Ph.D. School "Leonardo da Vinci" of the University of Pisa and works as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. Alessia is currently involved in EC funded projects for the aggregation, curation and export of library, archival and museum digital objects and metadata records. Her research interests include Digital Library Management Systems, data interoperability, compound object management, and service-oriented infrastructures. Email: alessia.bardi@isti.cnr.it

Candela, Leonardo
Leonardo Candela is a researcher at Networked Multimedia Information Systems (NeMIS) Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR - ISTI). He graduated in Computer Science in 2001 at University of Pisa and completed a PhD in Information Engineering in 2006 at University of Pisa. He joined the NeMIS Laboratory in 2001 and was involved in various EU-funded projects including CYCLADES, Open Archives Forum, DILIGENT, DRIVER, DELOS, D4Science, D4Science-II and DL.org. He was a member of the DELOS Reference Model Technical Committee and of the OAI-ORE Liaison Group. He is currently involved in the iMarine and EUBrazilOpenBio projects. His research interests include Digital Library [Management] Systems and Architectures, Digital Libraries Models, Data Infrastructures. Email: leonardo.candela@isti.cnr.it

Dušková, Marta
Marta Dušková studied library and information science at Comenius University in Bratislava (Slovakia). Since July 2010 she works in the Slovak Centre of Scientific and Technical Information in Bratislava (Slovakia) in Publication Evaluation Department. She deals with grey literature and The Central Registry of Publication Activity. She coordinates activities associated with obtaining and making grey literature available and cooperates with the processing and verification data publications included in The Central Registry of Publication Activity. She also participates as a team member at three national projects: National Information System Promoting Research and Development in Slovakia - Access to Electronic Information Resources, Infrastructure for R&D - Data Centre for Research and Development, and National Infrastructure for Technology Transfer in Slovakia. From 2012 she is studying PhD study at Comenius University in Bratislava (Slovakia) with the theme: Knowledge communication with support of grey literature. Email: marta.duskova@cvtisr.sk

Hardouveli, Despina
Despina Hardouveli is a senior Information Professional in Greece having an experience of nearly 30 years in the field. She graduated in Biology and holds a master degree. Since 1983 she works for National Documentation Centre (NDC/NHRF) in Greece. From 1993 to 2006 was Head of the S+T Information Services Department of NDC/NHRF, responsible for the provision of scientific & technological information to the Greek research and academic community. Since 1997, she has been a member of the project group for National Information System for Research and Technology (NIRST), a project elaborated within the NSRF (National Strategic Reference Framework) and the European Operational Program, responsible for several tasks, among them the settlement of international databases into the NDC's information systems, the support to the creation of the Electronic Reading Room, the transition from traditional retrieval services to digital content services, the conceptual design and content management of special bibliographic collections, the digital library integration. In charge of the creation of the NHRF's OA Repository "Helios", she is responsible for the content maintenance and upgrading, providing supportive services to the stakeholders/users. She was instrumental in the development of a funder repository of mainly grey literature material of diverse types, produced under the funding programmes of the Hellenic Ministry of Education (co-financed by the EU). She has served on numerous project committees and was demonstrator & trainer of the digital content services of NDC. From 2004 to 2010 was a member of GRNET S.A (Greek Research & Technology Network) Administrative Board. Email: dxardo@ekt.gr

Houssos, Nikos
Nikos Houssos works at the National Documentation Centre/NHRF in Athens, Greece as Head of the Software Development Unit. He is the software architect of the Greek "National Information System on Research and Technology" (EPSET) which comprises a variety of scholarly communications systems such as CRIS, repositories, e-publishing platforms and bibliographic systems. He has designed a number of grey literature systems including the Hellenic National Archive of Doctoral Dissertations. He participates in various EU FP7 projects like OpenAIRE/OpenAIREPlus, Arrow Plus, ENGAGE and PAERIP. He is a member of the euroCRIS Board and a contributor to the development of the CERIF data model. He has participated in EU FP5 and FP6 IST projects related to mobile networking and services (1999-2004), and lectured at the Technical University of Crete (2004-2007). He holds a Ph.D. in Computer Science from the University of Athens and has co-authored more than 30 peer-reviewed publications in international journals and conferences. nhoussos@ekt.gr

Kravjar, Július
Július Kravjar graduated in Mathematics at the Comenius University of Bratislava and later in Informatics. He is currently responsible for the „Central Repository of Theses and Dissertations" and „Plagiarism Detection System for Slovak Academic and Research Institutions" projects at the Slovak Centre of Scientific and Technical Information (SCSTI). Both nationwide systems are in the real operation from April 2010. He also participates as a team member at three other national projects: National Information System Promoting Research and Development in Slovakia - Access to Electronic Information Resources, Infrastructure for R&D - Data Centre

for Research and Development, and National Infrastructure for Technology Transfer in Slovakia. Prior to joining the SCSTI, he held several positions in software development and software and ICT services marketing in the private sector. julius.kravjar@cvtisr.sk

La Bruzzo, Sandro
Sandro La Bruzzo received his MSc in Information Technologies in the year 2010 at the University of Pisa, Italy. He is working as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. His research interests include Digital Library Management Systems, compound object management, and Service-oriented Data Infrastructures. Email: sandro.labruzzo@isti.cnr.it

Lin, Yongtao
Yongtao Lin has been working as a health information network librarian at the Tom Baker Cancer Knowledge Centre in Calgary Alberta since 2008. She provides library services to support health care professionals in their evidence-based practice. The library is part of a provincial patient-centered education strategy supporting cancer patients and their families. She was a hospital librarian in rural Nova Scotia for a few years before moving to the University of Calgary. Her prior experience as an instructor has led her to integrate education into various aspects of library programs. Yongtao is interested in the impact of grey literature in health care and a strong believer in evidence-based practice. Yongtao was the awarded the Canadian Hospital Librarian of the Year in 2011. yolin@ucalgary.ca

Manghi, Paolo
Paolo Manghi received his PhD in the year 2002 from the Dipartimento di Informatica of the University of Pisa, Italy. He is presently working as a researcher at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. His research interest include Data Models for Digital Library Management Systems, Types for Compound Objects, data curation in Digital Libraries, service-oriented ICT infrastructures with special focus on data ICT infrastructures. Paolo.Manghi@ISTI.CNR.IT

Nanakos Chrysostomos
Chrysostomos Nanakos has extensive experience in Open Source technologies. He has successfully executed several projects with different complexities and sizes in the Public and Private sector based on Open Source architectures. Since 2011 he cooperates with the National Documentation Center as a Senior Systems Administrator and Open Source Developer. He received his diploma in Electrical and Computer Engineering from the National Technical University of Athens (NTUA) where he is currently working towards the Ph.D. degree in Computational and Applied Electromagnetics. Email: cnanakos@ekt.gr

Pagano, Pasquale
Pasquale Pagano is a Senior Researcher at the Networked Multimedia Information Systems Laboratory of the "Istituto di Scienza e Tecnologie della Informazione A. Faedo" (ISTI) of the Italian National Research Council (CNR). He received my M.Sc. in Information Systems Technologies from the Department of Computer Science of the University of Pisa (1998), and the Ph.D. degree in Information Engineering from the Department of Information Engineering: Electronics, Information Theory, Telecommunications of the same university (2006). The aim of his research is the study and experimentation of models, methodologies and techniques for the design and development of distributed virtual research environments (VREs) which require the handling of heterogeneous resources provided by Grid and Cloud based e-Infrastructures. Pasquale has a strong background on distributed architectures. He participated to the design of the most relevant distributed systems and e-Infrastructure enabling middleware developed by ISTI - CNR. Pasquale is currently the Technical Director of the Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources (iMarine) and member of VENUS-C European project. He is also involved in the GRDI2020 expert working group a he serves EUBrazilOpenBio initiative as consultant. In the past, he has been involved in the D4Science-II, D4Science, Diligent, DRIVER, DRIVER II, BELIEF, BELIEF II, Scholnet, Cyclades, and ARCA European projects. Email: pasquale.pagano@isti.cnr.it

Picchi, Eugenio
Eugenio Picchi graduated in Computer Science, at Pisa University, is Research Director at the Institute of Computational Linguistics (ILC) of the National Research Council. He is currently responsible for the research line “Computational models and tools for research in humanities, with a special focus on linguistic and literary disciplines and on lexicography”. From 1972 to 1983 he was responsible of the Systems and Programming Division of the Linguistics Section of CNUCE (Pisa). Since 1983 he has been responsible of the Division “Methodologies and Tools for Lexicology and Computational Linguistics” of ILC. In the last few years he has been scientific director of national and international projects, among which: “International Network of Linguistics Data-Bases and Workstations” and “New Technologies for Language Engineering”, within the Project “Natural Languages Processing”; ILC Unit of Research of Esprit Basic Research Actions “Aquilex - Acquisition of Lexical Knowledge for Natural Language Processing Systems”, Action n. 3030; ILC Unit of Research of European Project MULTEXT “Multilingual Text Tools and Corpora”. He has also been Technical Manager of CNR in the European Project “EUROSEARCH” for an European Federation of multilingual WEB browsers and Scientific Director of the Project “Italian-Arabic bilingual Corpora and Tools” within the CLUSTER “Computational Linguistics: monolingual and multilingual Researches”, funding by Italian law 488/98. He has authored/co-authored a number of publications dealing with Computational Linguistics. eugenio.picchi@ilc.cnr.it

Roubani, Alexandra
Alexandra Roubani holds a Bachelor of Science (BSc), Technological Educational Institute of Athens, Department of Librarianship and Information Systems, with a Master’s Degree in Library Sciences (MLIS), Ionian University, Department of Archives and Library Science. She also studied Cultural Administration in the Department of Communication, Media and Culture, Panteion University of Social and Political Sciences (BSc). She is a cataloguing and metadata expert at Institutional Repositories of the National Documentation Centre (EKT)/National Hellenic Research Foundation since 1997, and her main occupation is focused on the gradual alteration/transformation of the National Union Catalogue of Serials. Her professional interests also include Authority Files, Digital Preservation Metadata Standards, Interlibrary Loan and Digital Curation. Email: arouba@ekt.gr

Sarantopoulou, Ioanna
Ioanna Sarantopoulou is the Head of the Digital Library Department of the National Documentation Centre (EKT) in the National Hellenic Research Foundation (NHRF). She is involved in organizing and maintaining Greek digital content in Library's collection development and also in providing intermediated library information services. She participated in several workgroups within the EKT for the implementation of the National Information System for Research and Technology. She formerly worked for many years at the EKT, using online systems for documentation and information retrieval, to facilitate information seeking for scientists on Natural Sciences and Engineering. She holds a Bachelor's degree from National Technical University of Athens in Chemical Engineering. Email: isaran@ekt.gr

Sassolini, Eva
Eva Sassolini graduated in Computer Science, at Pisa University, is CTER (Research Collaborator) at the Institute of Computational Linguistics (ILC) of the National Research Council - Pisa. She is involved currently in several national and international projects. Research Collaborator in "TextPower" (TP) project, (new technology and approach to treatment and exploitation of texts) and before in the project "Corpus Bilingue Italiano-arabo" for linguistic tools and resources for bilingual Italian/Arabic corpora realization. Junior researcher ILC in the project LINGUISTIC MINER: linguistic Knowledge system for the Italian language; working contribution in "INTERA" (Integrated European Language Data Repository Area) project, for multilanguage terms extraction. Junior researcher ILC in the project: "Progetto Iraq: navigazione nei materiali testuali del museo virtuale di Baghdad in maniera contrastava tra le tre lingue previste dal progetto (italiano, arabo e inglese)". Junior researcher ILC in the project 8: "Diffusione della cultura e valorizzazione del patrimonio letterario della lingua italiana e della lingua araba attraverso una diffusione telematica di banche di dati letterarie". Collaboration with IMSS (Istituto e Museo di Storia della Scienza) for the realisation of web applications for the query on galileian texts. Junior researcher in the project: "Corpus Bilingue italiano-arabo": in the framework of the comprehensive "Linguistica Computazionale: ricerche monolingui e multilingui". Email: eva.sassolini@ilc.cnr.it

Soumplis, Alexandros
Alexandros Soumplis studied "Engineering in Informational and Communication Systems" and holds a "MSc in Network & Data Communication" from Kingston University UK. Since 2010 is accepted by the Faculty of Sciences of the Hellenic Open University as a PhD student. His research interests involve informal learning environments as well as the use of innovative technologies for learning. Furthermore he has more than 12 years of active experience as a systems engineer with focus on core IT systems and in the past has worked for major computer and telecommunications companies in Greece. Since 2007 collaborates with the National Documentation Centre as a member of the "Information Systems and Networks Department" and has active involvement in several projects. Also, through his academic and professional career has submitted work and participated in several scientific and business conferences. Email: soumplis@ekt.gr

Stathopoulos, Panagiotis
Panagiotis Stathopoulos received his diploma in Electrical and Computer Engineering and his PhD in Broadband Networks at 1999 and 2004 respectively, from the National Technical University of Athens (NTUA). From 1999 until 2006 he has been with the Computer Networks Laboratory of NTUA participating and technically coordinating research projects in the areas of broadband communications and applications. From 2006, he is the head of the Systems and Networks Unit of EKT leading the team developing a highly sophisticated IT infrastructure, for providing advanced open access applications and services. He has taught at the University of the Aegean and the TEI of Piraeus, and he has over 30 publications in peer reviewed journals and conferences. Email: pstath@ekt.gr

Stathopoulou, Ioanna-Ourania
Ioanna-Ourania Stathopoulou received a B.Sc. in Computer Science and, later on, a Ph.D. degree from the University of Piraeus, Piraeus, Greece. Her doctoral research was sponsored by the General Secretary of Research and Technology of the Greek Ministry of Development, under the auspices of the PENED-2003 basic research program. Since May 2007, she is with the Department of Software Application Development of the National Documentation Centre (EKT) of the Hellenic Research Foundation, where she works as a software engineer participating in the design and implementation of digital repositories, e-publishing platforms and bibliographic systems. Her primary research interests are in the areas of affective computing, human-computer interaction, computer vision, and pattern recognition, and their applications in user modeling, information retrieval and intelligent software systems. She has over 30 publications in peer reviewed journals and conferences and has co-authored a monograph entitled "Visual Affect Recognition", published by IOS Press. Email: iostath@ekt.gr

Vaska, Marcus
Marcus Vaska is a librarian with the Health Information Network Calgary, Holy Cross Site, providing research and information support at an Alberta Cancer Care research facility. A firm supporter of embedded librarianship, Marcus engages himself in numerous activities, including instruction and research consultation, with research teams at Holy Cross. Marcus' current interests focus on educational techniques aimed at creating greater awareness and bringing grey literature to the forefront in the medical community. Email: mmvaska@ucalgary.ca

White, Gerry
Dr Gerald (Gerry) White is a Principal Research Fellow at the Australian Council for Educational Research (ACER). He specialises in the use of digital technologies and digital media in education and currently manages the Digital Education Research Network (DERN) (http://www.dern.org) which publishes weekly research reviews about ICT in education. Formerly head of Australia's education technology national agency for education and training, Gerry's interests and experience are in digital diffusion, online collaboration, grey literature, teaching and learning, leadership and online communities. Email: whiteg@acer.edu.au

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.

https://doi.org/10.26069/greynet-2026-000.548-tgj
TGJ Volume 9, Number 1 2013

An International Journal on Grey Literature

'ADAPTING NEW TECHNOLOGIES FOR GREY LITERATURE'

SPRING 2013 – TGJ VOLUME 9, NUMBER 1