https://doi.org/10.26069/greynet-2026-000.559-tgj

ISSN 1574-1796

An International Journal on

Grey Literature

Autumn 2010 – TGJ Volume 6, Number 3

'Research on Grey Literature in Europe'

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 6, Number 3, Autumn 2010. - *The British Library, FLICC-FEDLINK, INIST-CNRS, NTK, and NYAM are corporate authors and associate members of GreyNet International. This serial publication appears three times a year - in spring, Autumn, 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 grey literature community. It crosses continents, disciplines, and sectors both public and private. The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.

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

About *GreyNet*

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

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

Contents
"RESEARCH ON GREY LITERATURE IN EUROPE"

https://doi.org/10.26069/greynet-2026-000.559-tgj

EDITOR'S NOTE

As guest editor for this issue entitled "Research on Grey Literature in Europe", I have brought together six articles from authors and researchers in the Czech Republic and Italy. While Italy is well known for its contributions to this field of information dating back even before the First International Conference on Grey Literature in 1993, the Czech Republic has only recently focused on research in this field. To a great extent, this has been motivated by a project organized and managed by the National Technical Library in Prague. The project is discussed in the opening article of this journal.

Even more recently, we find that Czech research on grey literature has become aware of the advantages of collaborating with other researchers in Europe and abroad. During the Twelfth International Conference on Grey Literature in Prague, the results of one such study carried out with a Canadian researcher will be presented. The study deals with document typology, and the analysis is based on the classification of a number of systems including GreyNet, OpenSIGLE, Czech ASEP, RIV, et cetera.

In analyzing lists of document types, it was discovered that these typologies contain other aspects, besides "real" document types (reports, theses, etc.). Take for example events (arrangement, organization), types of events (conferences, speeches), producers (universities, institutes), processes (translations, output), content (political documents, legal texts), location (domestic, foreign), and format (e-texts, numeric data). However, since this approach is not systematic, it was decided to create a classification scheme for document types only, and then classify the other aspects into various groups in order to define them more precisely. The scheme will appear in text version and will schematically incorporate the use of mind maps.

It is believed that identifying a specific typology for credible grey literature document types, particularly reports, conference proceedings, and government documents, will assist in the classification of grey literature in the fields of science, research, and education. On the other hand, grey literature also consists of various means of communication, such as telephone calls, meetings, e-mails, blogs, interviews, social networking tools, and discussions in Wiki. And, these now present an added challenge for the grey literature community.

Petra Pejšová, Guest Editor

Grey Literature "information produced and distributed 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." (Luxembourg, 1997 expanded in New York, 2004)

https://doi.org/10.26069/greynet-2026-000.559-tgj

Czech National Repository of Grey Literature *

Petra Pejšová *(Czech Republic)*

Introduction

The National Technical Library (NTK) is a centralized professional library governed by the Ministry of Education, Youth and Sports in the Czech Republic. According to its statutes NTK is tasked with the management of a project, which deals with the construction of a National Repository of Grey Literature. This project is further supported by the Ministry of Culture and involves both a Functional Model as well as the Pilot Implementation of a Digital Library for Grey Literature. The project spans a four year period, which commenced in 2008 and will be completed in 2011. There are two participating institutions in the project both located in Prague - namely the National Technical Library and the University of Economics. In the past, NTK was responsible for the distribution of grey literature data from the Czech Republic to the SIGLE (System for Information on Grey Literature in Europe) database. This database was produced and managed by EAGLE, the European Association for Grey Literature Exploitation.

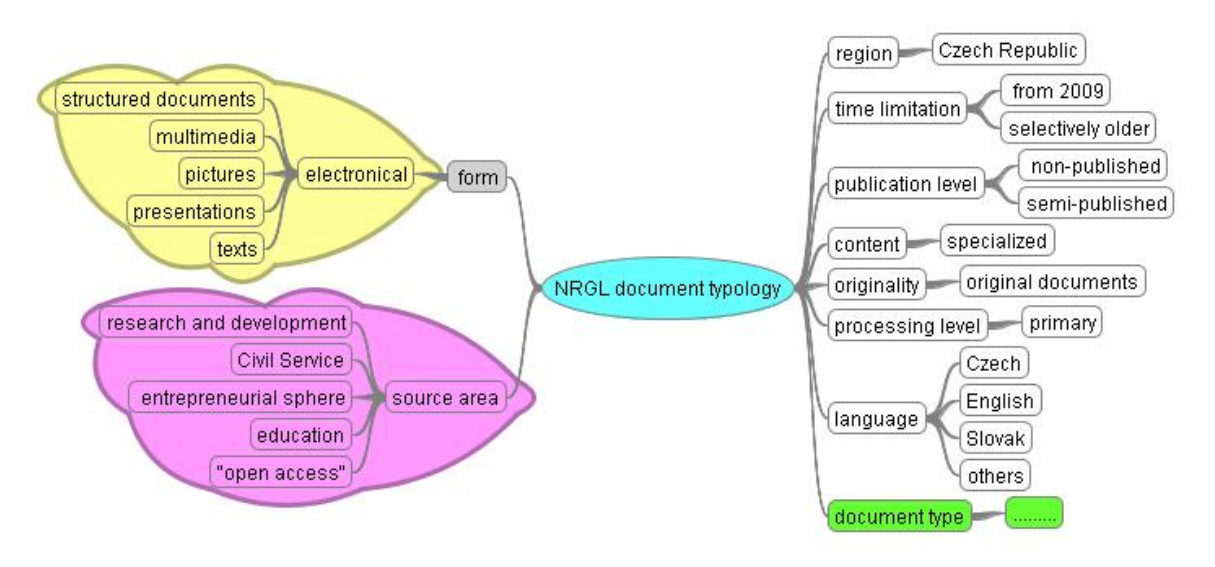
In 2005, the idea of a National Repository of Grey Literature first originated. The initiative was taken when both the input into the SIGLE database was terminated along with its concomitant cooperative system for grey literature in the Czech Republic. The project provides a working pilot application forming the basis of the National Repository of Grey Literature (NRGL). The project's standards and recommendations incorporate rules and methods related to metadata and interchange formats, online templates, license agreements, and other relevant issues dealing with grey literature. The standards and recommendations are to be published both in the Czech and English languages. Furthermore, the project places emphasis on building a network of cooperative organizations to support the coverage of grey literature in the Czech Republic. By way of a recent survey, the project was successful in establishing working ties with the Academy of Sciences Library, the Association of Czech University Libraries, and its Electronic Theses and Dissertations Working Group.

Typology

Before collecting grey literature using the NRGL system, a typology had to be defined to clearly specify the subject of collection. Most typologies, focus solely on the document type. For NRGL, a specific typology was created that also took other criteria into account. The NRGL typology is structured on two levels, to better capture individual aspects. The basic typology (level 1), depicted in the mind map (Figure 1), describes all document criteria adopted by NRGL. The basic criterion consists in the collection of documents in electronic form (i.e., documents originally created in digital form or subsequently digitized) since 2009, with older documents processed on a selective basis. Metadata on all accessible grey literature records in the Czech Republic is collected without restriction. Grey literature is supplied by partner organizations in the area of education, research and development, Civil Service, the business sphere and verified open access sources. Grey literature collected in the NRGL system must have originated in the Czech Republic and its contents must be original and specialized. Preferred languages for items are Czech, English, Slovak and other European languages on a selective basis.

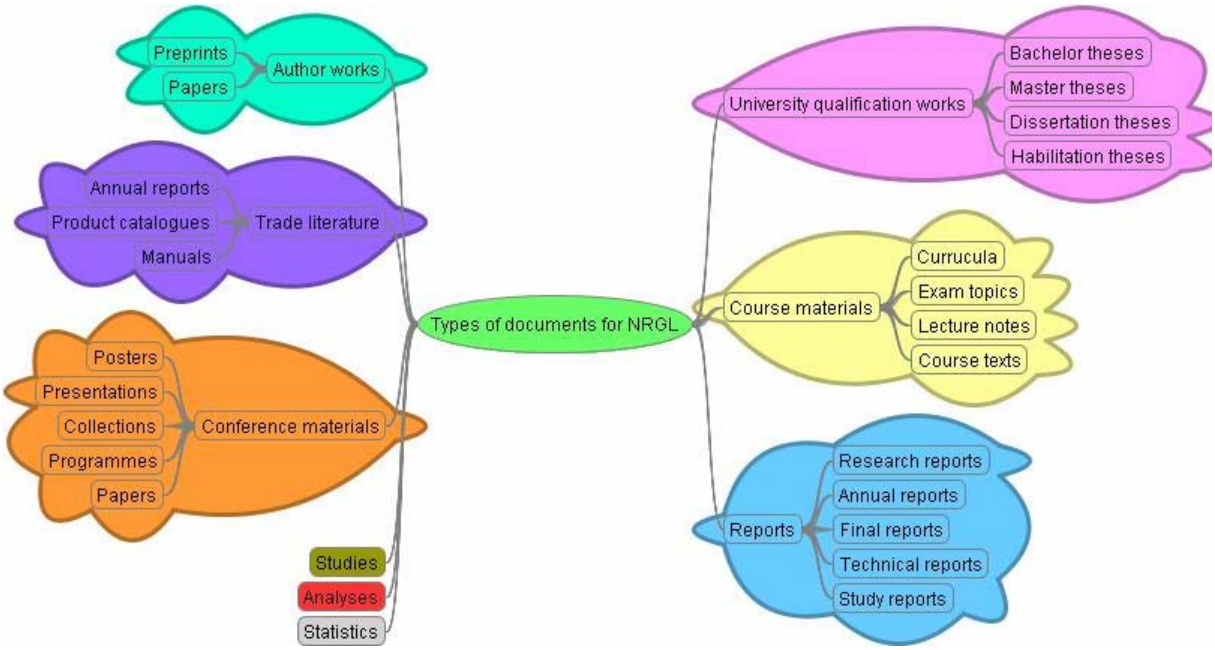
* Based on a paper presented in 2009 at a seminar entitled *'On Providing Access to Grey Literature'*, Brno (CZ)

Figure 1: Basic NRGL typology (level 1).



Level 2 of the NRGL typology, concerned with document types, is based upon level 1. The main document groups consist of ETDs, course materials, work by authors, corporate literature, conference materials, etc. Individual groups contain specific types of documents. Another project aim is to verify which types of documents are more numerous, and to extend the typology by adding new document types needed for science and education.

Figure 2: Typology according to document types (level 2)



Software Solution

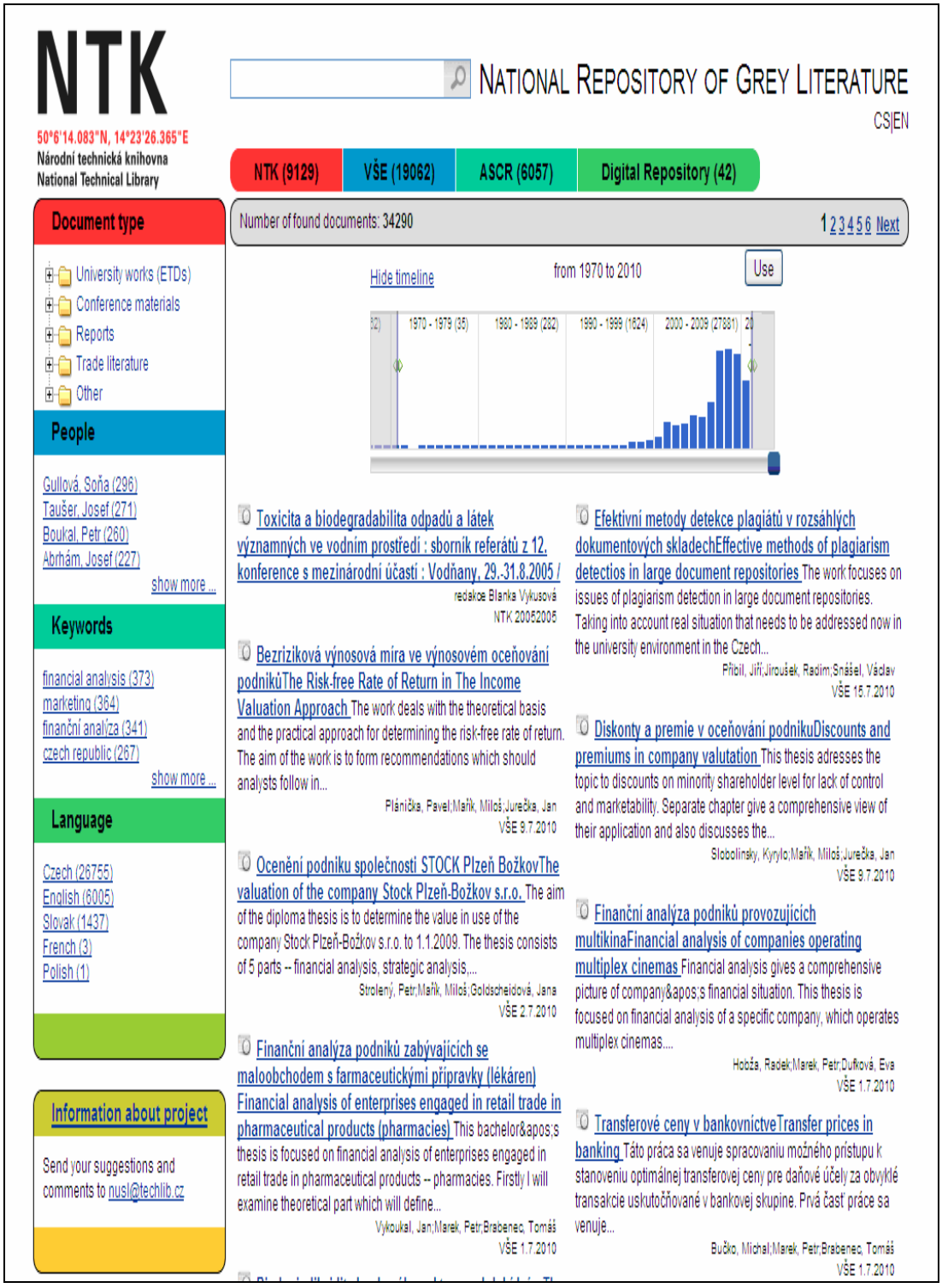
The software solution for the NRGL is composed of two parts - the CDS Invenio[1] system for digital repositories and indexing, and the FAST ESP[2] search system for the user interface. The solution led to a user interface being constructed for the digital repository which enabled easy, friendly search and data access, as well as functionality integrating the platforms of independent repositories. The same architecture has been used successfully for several years at CERN in Switzerland.

CDS Invenio is an open source software licensed under the GNU GPL, developed by the CERN organization and EPFL (École Polytechnique Fédérale de Lausanne) for general administration of documents, institutional repositories, and/or large-size library systems. CDS Invenio is free for installation, usage and adaptation. CDS Invenio has been localized to eighteen (18) languages including the Czech and Slovak languages. The CDS Invenio system is highly structured and flexible, and it can be customized in several ways for various uses. It enables the user to define ones own metadata schema, having the internal naming convention of MARC21 as its underlying standard. It likewise supports the OAI-PHM standard.
Furthermore, it is important to know that CDS Invenio is possible to install on GNU/Unix system, it uses an MySQL database server and an Apache/Python web server. Apart from the system itself, we can also expect that some of the services are localized to Czech or Slovak needs and requirements. Access to NRGL via CDS Invenio is available at http://invenio.ntkcz.cz.

Figure 3: The input screen of CDS Invenio

The FAST ESP indexing and search system provides user-friendly, easy data access. The FAST ESP interface provides a secure, scalable and relevant search, including declension of Czech terms. The characteristics of the FAST ESP solution make it possible to identify the context and purpose of the query, find the corresponding terms both in the metadata in the documents and return a response in context. Users benefit from precise results and contextual dynamic navigation enabling them to further search for related information. The FAST ESP is used as a central user interface and is operational since late 2009. All institutions in the NRGL network are gradually integrated into this interface, which is available at www.nusl.cz .

Figure 4: NRGL user interface



Support and Cooperation

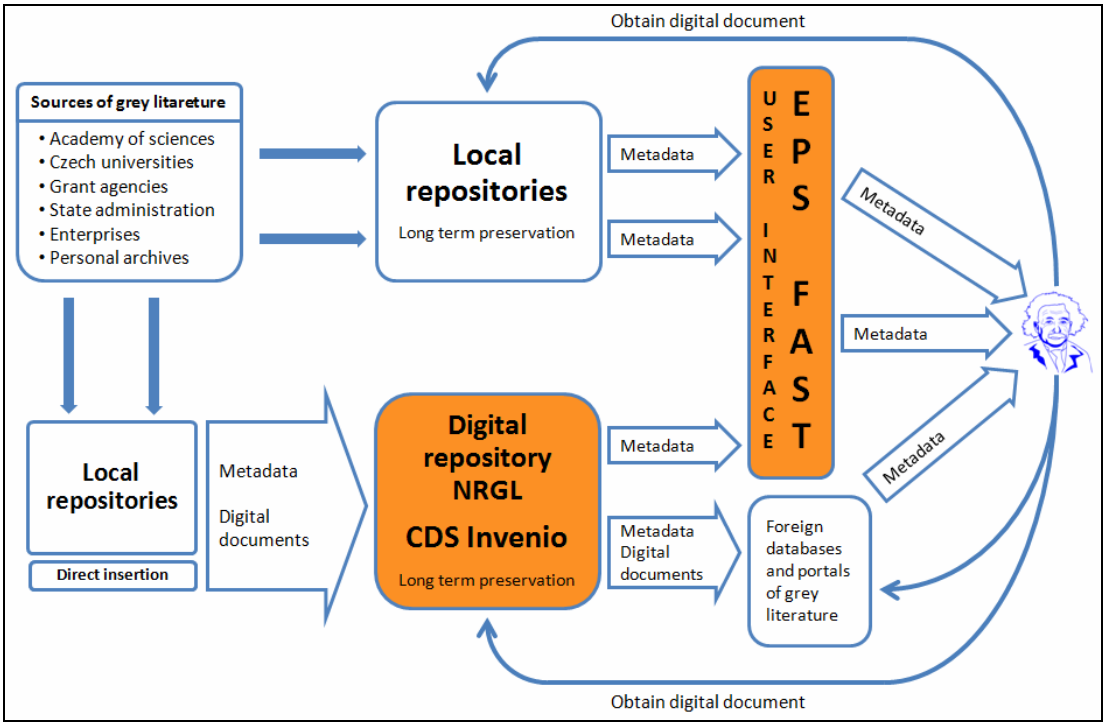
The acquisition of grey literature in the Czech Republic is aimed at the systematic collection, long-term archival and provision of access to grey literature for users in science, research and education. The credibility of information thus acquired is considered paramount. This is ensured by cooperation with credible producers of grey literature. As a result, unverified grey literature such as blogs, phone calls, discussions, and e-mails are not subject to acquisition.

Legal expertise for NRGL was acquired by focusing on the potential for data acquisition, storage and accessing grey literature in the Czech Republic. This legal expertise holds for

both data producers as well as database administrators. A special part of the repository is deals with dissertations, because this kind of grey literature has another legal framework than the rest of grey literature in the Czech Republic. In the spirit of collaboration with Czech universities, part of the legal expertise aims at recommendations to universities on how to facilitate license agreements with students. And, altogether with this legal expertise we were able to draft examples of license agreements for further collaboration with data producers and administrators.

Technical collaboration with data producers is possible in four ways. Firstly, by directly joining a local database using ESP FAST to a central user interface. In this case it is necessary that data producers have local databases available through OAI-PMH protocol. Here, data are not deposited and archived in the NRGL digital repository (CDS Invenio in NTK), but instead are available only through the NRGL central user interface. Secondly, if data producers seek to harvest their local database via the NRGL digital repository, it is initially required that they deposit and archive their data in NRGL. Thirdly, by using CDS INVENIO as a local database distributed by NTK in a default version and harvest it as other local bases in the NRGL digital repository. And finally, by entering records and digital documents directly into the NRGL digital repository. This option is highly recommended for data producers that have low or minimum data volume. All four scenarios are geared to a central user interface available at http://www.nusl.cz.

Figure 5: NRGL cooperative scheme



Project Websites

In 2009, the project's websites also became available in English. The English language version follows the Czech original. These websites feature continuously updated information on the NRGL project and its outcomes. Furthermore, these websites support collaboration between NRGL and grey literature producers, for example in matters dealing with legislation that touch on grey literature. They also offer a downloadable software solution for local installation of a digital repository. And, in order to enable broader connections, selected grey literature information sources are cited, along with references to portals supporting research and development in the Czech Republic.

The website may be found at http://nrgl.techlib.cz/. The Media wiki application and the WordPress graphical style modified to follow NTK graphical guidelines were used in creating the site. The media wiki application supports participation in content creation within the working group and the opening of discussion groups on various topics for participation by the wider public.

Figure 6: Screenshot of the project homepage

Workshops

Since 2008, an annual *Seminar for the Access of Grey Literature* has been organized to inform and support scientific discussion concerning grey literature. During the seminar, updated results of the NRGL project are presented and topics relating to grey literature are discussed such as systems dealing with grey literature, formats and standards for long-term archival, standards for describing and exchanging sources, and copyright issues. All presentations and the full text of lectures are available in the online proceedings via the project website. Presentations during the seminars are both in the Czech and English language; however, the proceedings appear only in English. The reason for this is to reach and inform the international grey literature community on developments in the Czech Republic. The third Seminar in this series will be held on December 8th 2010, immediately following the Twelfth International Conference on Grey Literature at NTK in Prague on December 6-7, 2010

For more information and links NRGL in English:
The project webpage: http://nrgl.techlib.cz
User interface: www.nusl.cz (select English)
Digital repository: http://invenio.ntkcz.cz (select English)
The project contact e-mail: nusl@techlib.cz

References

[1] http://cdsware.cern.ch/invenio/index.html
[2] http://download.microsoft.com/download/1/4/8/1483939B-15B8-4DD3-B06D-204D03EC8A1E/Fast_ESP_Prod_Guide.pdf

Licensing the Use of Grey Literature under the Czech Law*

Radim Polčák *(Czech Republic)*

Introduction

The concept of grey literature obviously does not have legal nature. It emerged as factual consequence to growing production of various kinds of literary works whose primary aims and applications somehow differed from conventional literature. However, when Gokhale[1] tried to precisely define and categorize grey literature, it turned out that this general category breaks down to dozens of subtypes and subcategories. Moreover, it seems that the term "Grey Literature" might even have different meanings for different areas of knowledge[2]. As the broadness and complicated structure of grey literature makes it difficult even for a librarian to precisely define what does the term in fact mean, it is then even more difficult for a lawyer to add its precise legal understanding. Therefore, we need first to clarify at least basic legally established criteria for the distinction between grey and other 'coloured' literature in order to be able to define its legal regimes and subsequent rights and duties.

As pointed out above, grey literature is to be generally understood as having slightly different basic nature. When trying to find out some general point of distinction, we can start with very general and not entirely precise but relatively simply conclusion that the basic aim of grey literature differs by the fact that its primary aim is not to be publicly issued and disseminated. While the reason for the creation of (general) literature is to broadly disseminate the ideas of its creators, the reason for the creation of grey literature is mostly to be seen in:

1. Fulfilling the conditions of various study programmes (namely at universities)
2. Reporting activities of academic institutions incl. the progress or outcomes of research projects
3. Confronting ideas within limited groups of specialists (namely at conferences or in the process of drafting various publications)
4. Defining technical standards
5. Fulfilling legal obligations (namely in bookkeeping and official reporting)

Grey and general literatures differ also in their availability. As the main aim of grey literature is other than to be (truly) published[3], its general public availability is relatively low. Then, there are lots of publications falling within the category of grey literature that are on one hand highly scientifically valuable, useful or inspiring, but on the other hand difficult to be spotted and obtained. Consequently, we see significant development of various activities aiming broader availability of grey literature including its classification, systematic storage and public dissemination. Neither in the Czech Republic, nor in the rest of the world has the presence and significance of grey literature been directly acknowledged by black-letter law. Thus, there is no special legal definition or concept that would precisely reflect the existence of literary works of the above indicated different teleology. However, it does not mean that it would be impossible to legally handle grey literature, i.e. to speak about rights arising of the process of its creation, storage and use.

Grey literature in the Czech Law

In the following paragraphs, we will briefly focus on the forms in which Czech law applies namely on legal issue of gray literature that we see as crucially important, i.e. on its legal licensing. It should be noted in advance that the concept of copyright in the Czech Republic is no different from other European states, as Czechoslovakia and then Czech Republic were members of basic harmonizing international conventions such as:

- Convention establishing the World Intellectual Property Organization (WIPO/OMPI) - a member since 1970
- Berne Convention for the Protection of Literary and Artistic Works - a member since 1921.
- The Agreement on Trade-Related Aspects of Intellectual Property Rights - a member since 1995.
- WIPO Copyright Treaty - a member since 2002.

* Based on a paper presented in 2009 at a seminar entitled 'On Providing Access to Grey Literature', Brno (CZ)

In the process of the accession of the Czech Republic to the EU, the copyright provisions of the Czech law were harmonized also with EC directives such as:

- Council Directive 92/100/EEC of 19 November 1992 on rental right and lending right and on certain rights related to copyright in the field of intellectual property
- Council Directive 93/83/EEC of 27 September 1993 on the coordination of certain rules concerning copyright and rights related to copyright applicable to satellite broadcasting and cable retransmission
- Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights
- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases
- Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society
- Directive 2004/48/EC of the European Parliament and of the Council of 29 April on the enforcement of intellectual property rights

Similarly as in the case of other European countries, the concept of Czech copyright law is based on a distinction of moral and material rights of an author[4]. The moral rights include namely:

- authorship right - a right to be indicated as an author
- integrity right – a right for the preservation of the integrity of the work (incl. exclusive right to permit modifications)
- publication right – a right to decide upon first publication of the work[5]

The moral rights belong exclusively to the author and, importantly for legal regulation of grey literature, they represent the only copyrights that might be executed by a third person even upon one-sided permission of the author. On the contrary, the material rights represented by the most important rights to use the literary works, can be executed by anyone else than the author only upon an agreement. Taking the point of view of the creation, processing and dissemination of grey literature, we shall focus namely on material rights, i.e. on the rights of usage of the respective literary works.
In the case of regular publications, the author concludes a contract with a publisher upon which the respective literary work is processed and exclusively disseminated by the publisher. As noted above, one of the differences between regular and grey literature is to be seen in the fact that grey literature is regularly not created primarily in order to be published. Moreover, in most cases of grey literature, the publisher is entirely missing. This means that when analyzing subsequent legal relations arising of the creation of grey literature[6], we have to take into account namely specific forms in which material copyrights are licensed. Consequently to our basic classification of general purposes of grey literature, we can distinguish the following legal regimes:

1. Fulfilling the conditions of various study programmes – works created by students in the course of their studies
2. Reporting activities of academic institutions incl. the progress or outcomes of research projects – works created by researchers directly for their research institutions or projects
3. Confronting ideas within limited groups of specialists (namely at conferences or in the process of drafting various publications) – works created by independent individuals and published as manuscripts or articles in conference proceedings
4. Defining technical standards – works created by employees of state, unions, guilds or similar organizations
5. Fulfilling legal obligations – works created by employees of business units or state organizations

From the point of view of the Czech copyright law, the above types of grey literature can be subsumed under the following categories:

A. Falling outside of the scope of copyright law – those of sub 4. and sub 5. that are either not of creative nature (accounting books, filled forms etc.) or that have a form of official publications (texts of laws, decrees, judgments, etc.)
B. Student works – namely sub 1.
C. Employment works – namely sub 2. and those of sub 4. and sub 5. that do fall within the scope of copyright.
D. Contracted works – namely those of sub 2. in the case when a report is not produced by employees of the respective institution but upon a contract by 3rd person (for example by project management agency)
E. Individual works – namely sub 3.

As the category sub A. falls outside of the scope of copyright protection, it has no generally definable legal regime[7]. The category sub E. represents from the legal point of view normal (unspecific) creation that is being handled by the copyright law as any other literary work. The rest of categories include works with specific legal regimes that shall be briefly described in particular.

Student works in grey literature

As noted above, this category includes namely works that are created in the course of studies. It includes various dissertations, seminary works and other types of works that are used namely in order to prove the fulfillment of duties prescribed for the respective types of studies.

These works are authored by students. However, the respective schools also hold some rights that reflect the fact that the creation of the works was sponsored by various means (consultations by staff members, library resources etc.). The Czech law defines the rights arising of the creation of student works in Article 60 of the Copyright Act[8] that read as follows (informal translation):

(1) A school or school or educational establishment shall have the right to conclude, under habitual terms, a licence agreement on the utilization of a school work (Article 35 paragraph (3)). Where the author of such work has refused to grant his permission without stating a serious reason, such entities may claim compensation for the absence of manifestation of will in court. The provision of Article 35 paragraph (3) shall remain unaffected.
(2) Unless agreed otherwise, the author of a school work may use his work or grant the licence to another party if this is not in contravention of the legitimate interests of the school or school or educational establishment.
(3) The school or school or educational establishment shall be entitled to claim from the author of the school work, from the income earned by him in connection with the utilization of the work or granting of the licence pursuant to paragraph (2), an appropriate contribution to the reimbursement of the cost incurred by them due to the creation of the work, and that, depending on the circumstances, up to the full amount of this cost; the determination of the amount shall take into account the proceeds from the utilisation of the school work pursuant to paragraph (1) received by the school or school or educational establishment.

As implied from the above, the schools might use the works of their students freely for internal purposes[9]. In the case the schools would like to use these works in any other manner (for example to publish them or to upload them to databases of grey literature operated by a 3rd person), there is a need of a conclusion of a license agreement. There is, however important exception for works that fall within the scope of the Higher Education Act[10], i.e. bachelor theses, master theses, rigorous theses and doctoral dissertations. In the case of these theses, the schools have a duty (and obviously the right at the same time) to make them available to the public. This duty that includes operating publicly accessible database arises of Article 47b(1) that reads as follows:

(1) Higher education institutions are obliged to make public, at no profit to themselves, the doctoral, Master's, Bachelor's and advanced Master's ("rigorózní") theses that have been defended at their institutions, including the readers' reports and results of the defence. The institution will do this by making available a database of these theses. The means of providing access to these theses is stipulated in the internal regulations of the higher education institution.

The above article gives higher education institutions, namely universities, a right of creating and operating databases that are used in order to process the above listed types of student works. Schools are in this case in the position of a maker of a database what gives them

rights not just to make the database publicly available but also to non-commercially dispose with it as defined in Article 90(1-3) of the Copyright Act as follows:

(1) The maker of the database shall have the right to extraction or re-utilisation of the whole content of the database or of its qualitatively or quantitatively substantial part, and the right to grant to another person the authorisation to execute such right.
(2) The extraction pursuant to paragraph (1) means the permanent or temporary transfer of all or substantial part of the content of a database to another medium by any means or in any form.
(3) The re-utilization pursuant to paragraph (1) shall mean any form of making available to the public all or a substantial part of the content of a database by the distribution of copies, by rental, by on-line or other forms of transmission.

It should be noted that schools are entitled only for non-commercial disposal of the above database rights. Thus, in the case when a school decides to cooperate with an operator of grey literature database, the licenses might permit only non-commercial reuse of the content. Moreover, the above implied license for operating databases of school works does not include all types of school works , but only those listed under Article 47b(1) of the Higher Education Act. Thus, the publication or re-use of any other works require having prior license agreement with their authors (i.e. the students).

Employment works in grey literature
As noted above, substantial proportion of grey literature has the form of annual reports, research reports or other documents that are created in order to record the activities of various academic institutions like universities or research institutes. These documents are created as a byproduct in the course of research or other academic activity and are authored mostly by academic and/or administrative staff.
Czech legal regime of employment works is relatively simple. Although there are no exceptions as to the attribution of moral or material rights to the respective authors, both of these rights are exclusively exercised by the employer as laid down in Article 58 of the Copyright Act:

(1) Unless agreed otherwise, the author's economic rights to a work created by the author in fulfilling his duties araising from the employment or civil service contract to the employer or from an employment relationship between a cooperative and its member (henceforth referred to as employee work) shall be exercised exclusively by the employer in his own name and on his own account. The employer may assign the exercise of the right pursuant to this paragraph to a third person only with the author's consent, unless this occurs during the sale of the undertaking or of its part.
(2) In the event of the death or dissolution of the employer who has been authorised to exercise the economic rights to an employee work, and who has no legal successor, the authorisation to the exercise of these rights shall fall to the author.
(3) Where the employer does not at all utilize the economic rights to an employee work, or utilizes them inadequately, the author shall have the right to ask the employer to grant him the licence under habitual conditions, unless there is a serious reason on the part of the employer to refuse it.
(4) The author's personal rights to an employee work shall remain unaffected. Where the employer exercises the economic rights to an employee work it shall be deemed, however, that the author has given his consent to the work being made public, altered, adapted including translation, joined with another work, included into a collective work and, unless agreed otherwise, also being introduced in public under the employer's name.
(5) Unless agreed otherwise, it shall be deemed that the author has given the employer his consent to complete his unfinished employee work in the case where it has not been finished in spite of a challenge by the employer to complete the employee work within an extended period, or when the employee's obligation to complete such work has expired due to his death or due to the impossibility to meet this obligation.
(6) Unless agreed otherwise, the author of the employee work is entitled to an appropriate supplementary remuneration from the employer if the wages paid to the author by the employer are in evident disproportion to the profit from the utilisation of rights to the employee work and to the importance of such work for the achievement of this profit; this provision shall not apply to works referred to in paragraph (7), be they employee works or considered as such, unless agreed otherwise
(7) Computer programs and databases, and cartographic works which are not collective works shall be deemed employee works also where they have been created to order; the person who ordered them shall in such case be considered the employer. The provisions of Article 61 shall not apply to such works.

(8) The provisions of paragraphs (1) to (6) shall remain unaffected by the termination of the legal relationship pursuant to paragraph (1) or, respectively, paragraph (7).

The above implied license gives the employer almost total control over moral and personality rights to the copyrighted works of its employees. It means that the employer might indicate itself as the author and copyright holder of the respective document, it might freely use or license it and it might also freely update or modify it even after the termination of the labour contract. The employer is in this case also fully entitled to dispose with copyrights what enables it to contract with publishers or to make the documents available for any form of re-use including any form of transfer to a database of grey literature. Compared to implied legal position of schools in previously explained case of student works, employers dispose with copyrights not just to the whole database, but also to every single document. Moreover, the copyrights to employment works might be exploited by the employer also commercially.

Contracted works in grey literature

The only significant formal difference between employment works and contracted works is to be seen in the nature of the relation between authors and users of their works. While employment works are created upon labour or similar contracts, contracted works are produced upon business agreements. This basic difference is reflected by implied license in Article 61(1,2) of the Copyright Act that reads as follows:

(1) A work created on the basis of a contract for work (a work created to order) may be used by the person who ordered it only for the purpose defined by the contract. Unless stipulated otherwise by this Act, the customer shall be authorised to use the work in extension of such purpose only on the basis of a licence agreement.
(2) Unless agreed otherwise, the author may use the work made to order himself and to grant a licence also to another party if this does not contravene the legitimate interests of the customer.

As it can be seen, the above implied license includes very limited set of rights. However, if the object of above mentioned contract is a document of grey literature, we can formulate an opinion that its re-use naturally falls within the scope of the purpose of the contract as laid down in Article 61(1), first sentence. In any case, it is highly advisable in cases of contracted production of grey literature to pay enough attention to precise defining of licenses and eventual exclusivities, as the law leaves the field of licensing of such works almost entirely opened for mutual stipulations.

Concluding remarks

As grey literature shares almost in full the legal regime with other copyrighted literary works, any processing including the most common ways of its dissemination through academic databases has to be properly licensed. Unlike in the case of a dissemination of (regular) literature where licensing and publishing is relatively simple and straightforward, licensing and sublicensing of grey literature is in most cases multi-level and relatively complicated. Moreover, legal handling of grey literature in most cases relies not on contracted but rather on implied licenses.

All the above mentioned factors cause the legal regime of regular ways of creation and dissemination of grey literature to be less certain and more risky than in the case of regular publications. In that context, it is quite problematic to simply rely on an assumption that these risks will be constantly mitigated by the fact that there are still missing particular legal interests and consequent claims.

This brief note aimed on giving basic overview of specific legal regimes of implied licensing of most common works in grey literature under the Czech Law. As the Czech law, similarly to other European jurisdictions, does not reflect the existence of grey literature as specific concept of literary work, we see here significant risk in the development of international systems for the creation, processing and dissemination of grey literature without careful implementing local legal standards.

Thus, national specific features namely of implied licensing still represent major legal constraint to the development of cross-border databases of grey literature. As more particular harmonization of this aspect of copyright is not very likely to take place in foreseeable time, it is and will remain a must for designers of various systems processing grey literature to back them not just by proper (universal) concept, system and technology, but also by country-specific analysis and implementation of applicable legal regimes.

.

References

[1] See Gokhale, Pratibha. Grey Literature Varieties - Definitional Problems, in Third International Conference on Grey Literature : Perspectives on the Design and Transfer of Scientific and Technical Information, 13-14 November 1997 (Conference Proceedings), Amsterdam : GreyNet, 1998, pp. 259-273.

[2] This can be a consequence of the fact that different disciplines work with different scientific procedures, publication usages and, probably most importantly, with different business models.

[3] It is also typical that grey literature is being created or produced but not truly published, so that one of its typical features is the lack of the presence of a publisher in the production chain.

[4] This distinction originates in the Berne Convention for the Protection of Literary and Artistic Work

[5] Grey literature is often indicated also as 'unpublished' literature. If we take this term from the legal point of view, we have to state that in fact it cannot apply on most of documents of grey literature, as they have been made available to the public for example by publicly defending a thesis or by displaying a research report on the internet.

[6] Here, it should be noted that copyrights arise at the moment of the creation of the respective literary work.

[7] Any protection of this type of documents is based on their content. Thus, these documents might be particularly protected as know-how, trade secrets, personal data etc.

[8] The Act No. 121/2000 Sb., on Copyright, Rights Related to Copyright and on the Amendment of Certain Laws (Copyright Act). Unfortunately, there are no formal translations of Czech laws available and there is also no general database of translated Czech black-letter law. The informal translations used in this article are taken from WIPO database at http://www.wipo.int/clea/en/text_html.jsp?lang=en&id=962.

[9] One of contemporary most frequent form of use is the inclusion of these works into referential databases for fighting plagiarism.

[10] The Act No. 111/1998 Sb., on Higher Education Institutions and on Amendments and Supplements to Some Other Acts (The Higher Education Act). The informal translation used in this article was taken from the server of Masaryk University at http://www.muni.cz/general/legal_standards/higher_education_act.

https://doi.org/10.26069/greynet-2026-000.559-tgj

Polythematic Structured Subject Heading System & Creative Commons *

Jindřich Mynarz, Kateřina Kamrádková, and Kristýna Kožuchová *(Czech Republic)*

Abstract
This article presents how knowledge organisation systems in general and Polythematic Structured Subject Heading System in particular can be used in the context of grey literature resources. Starting from improved accessibility and findability, employing knowledge organisation systems can benefit both authors of grey literature and users. It suggests some of the functions knowledge organisation systems can fulfill in the grey literature environment, such as metadata authoring, search query refinement, or alternative ways of navigation. Polythematic Structured Subject Heading System (PSH) is an example of knowledge organisation system that was published under the Creative Commons licencing terms in the form of linked data. The article discusses some of the practical applications of PSH with hands-on examples. Second part of this text focuses on Creative Commons (CC) licences which provide publishing protocol particularly suited for the field of grey literature. CC offers an easy way to publish grey literature in such manner that it is more accessible and findable without any help from commercial publishers.

PSH terms
knowledge systems , subject information languages , grey literature , intellectual property law, licences

The focus point of this article is Polythematic Structured Subject Heading System (PSH), a subject heading system produced by the National Technical Library of the Czech Republic. As an example of knowledge organisation system, it has many useful applications in the field of grey literature. Within this context it can help both authors (publishers) and users (consumers) of grey literature resources. PSH has been published under Creative Commons licencing terms and it conforms with the *linked data* principles. These features, which are described in detail in the following text, greatly extend the range of possible uses of PSH, most of which are applicable in the area of grey literature that frames this article's scope.

1 Context of grey literature

According to the generally accepted definition, grey literature is *„the information 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"* (PERSPECTIVES, 1997).[1] If some resource is labeled "grey", it does not say anything about its content. *„Grey does not imply any qualification, it is merely a characterisation of the distribution mode"* (MACKENZIE OWEN, 1998). It is common though that the content of grey literature has some distinctive characteristics. It is by far the most immediate channel for new information stemming from the research, government or educational institutions. Also, it tends to be less permanent because it escapes the established means to preserve information (e.g. libraries).

Grey literature is an important source of information and it constitutes an alternative to the conventional white literature. It includes all types of literature that are not available through the traditional publishing channels. Especially the Internet has enabled dissemination of these publications that have not been formally published to a much wider audience.

As the definition of grey literature stands, it lacks "commercial control". Often it turns out that this feature also implies a lack of "bibliographic control". Frequently it is up to the creators of grey literature themselves to provide bibliographic information for their publication so that it can be indexed and found. *„And because there are no commercial issues at stake, improved referencing will always be dependent on the initiatives and willingness of producer or distributor organisations and a few committed professionals"* (SCHÖPFEL, 2006). From this issue arises the challenge of the quality of indexing and subject access. *„Self-posting is perhaps the most used method to disseminate information but it is not effective because of a lack of adequate indexing"* (DE BLAAIJ, 2005). Even the content of grey literature may *„represent topics that are too new or different to easily provide subject access for"* (CHILDRESS, 2003).

* Based on a paper presented in 2009 at a seminar entitled 'On Providing Access to Grey Literature', Brno (CZ)

2 Role of knowledge organisation systems

Institutional repositories or digital libraries where grey literature is often kept can alleviate specific issues of grey literature by the use of a knowledge organisation system (KOS), such as controlled vocabulary, thesaurus or subject classification. Employing knowledge organisation system has many advantages both for authors (publishers) and users. They act as metadata standards which ensure consistency of subject indexing that is often lacking in subject organisation based on free keywords. Knowledge organisation system *„attempts to increase consistency and improve access to digital collections and web navigation systems via vocabulary control"* (TUDHOPE, 2006). The general purpose of vocabulary control is to reduce the ambiguity of natural, uncontrolled language.

Functions addressing this purpose are grouped under the term terminology services. These include *„resolving search terms to controlled vocabulary, disambiguation services, offering browsing access, offering mapping between vocabularies, query expansion, query reformulation, combined search and browsing"* (TUDHOPE, 2006). The reduction of vagueness of natural language can be achieved by knowledge organisation system that defines the scope of a term, provides a set of synonyms or restricts the scope of a term (TUDHOPE, 2006). Free keywords are not present with these characteristics and thus searching with them might not yield relevant results. As Tudhope puts it, *„when searching free text with uncontrolled terms, significant differences can stem from trivial variations in search statements and from differing conceptualisations of an information need"* (TUDHOPE, 2006). Using controlled vocabulary provides a way to eliminate these variations by adhering to the terms representing precisely defined concepts which increases both precision and recall of the search query. KOS performs as a common ground which *„provides potential pathways (for human and machine) that connect a searcher and indexer's choice of terminology"* (TUDHOPE, 2006).

Metadata drawn from KOS can increase findability of grey resources because they can be crawled and indexed by harversters or semantic search engines (e.g., Sindice[2]).

Grey literature is often characterised by its limited availability. Also, the question of preservation is frequently left unresolved. These issues can be alleviated by the use of metadata, which can *„provide information required for permanent access to the digital object."* (HOORN, 2006). This implies the need of metadata standardization in the areas such as *„persistent identifiers, representation of vocabularies, protocols for programmatic access"* (TUDHOPE, 2006). Interoperability of KOS depends on a global identifier mechanism for referring to vocabulary concepts. From the proposed identifiers currently *„only "http"URIs offer a simple, widely deployed dereferencing mechanism"* (TUDHOPE, 2006). In this way, the necessary infrastructure is created upon which further services can be built.

2.1 Assisting authors

Employing KOS can provide advantages both for authors of grey literature and users. From the authors' point knowledge organisation systems can facilitate metadata authoring. Because of the self-published nature of grey literature it is often the case that the document metadata is made by authors themselves rather than by library professionals. In order to achieve reasonable indexing quality tools for metadata authoring can be utilised. The barrier of complicated metadata formats can be lowered by tools generating *metadata snippets* (short fragments of metadata) that can be readily embedded in documents (web pages, PDFs).

The process of subject terms assignment is considered difficult. The first step to make this easier is to simplify the selection of suitable indexing terms. Indexing system should have an easy-to-use interface that allows intuitive navigation through the knowledge organisation system (e.g., using visualization techniques), preferably with some helper tools for spelling correction (cf. "grey literature"and "gray literature"[3]) or term suggestion.

Next step further is automatic indexing that facilitates automatic subject term assignment according to the information content of the indexed document. It involves techniques of natural language analysis, subject analysis, machine learning processes, or simple phrase matching. The example of automatic indexing is term extraction technique aiming to extract from the document the terms and phrases that make the most accurate representation.

This approach is used in integrated digital library system CDS Invenio which will be used for the National Technical Library's *National Repository of Grey Literature project*[4]. CDS Invenio *BibClassify module*[5] accepts controlled vocabularies in Simple Knowledge Organisation System (SKOS) format. It analyzes the document provided by the user and then it produces a list of suggested terms from the supplied knowledge organisation system recommended to be used in the document metadata. However, it does not use any machine learning or artificial intelligence techniques and therefore it produces only approximate results.

For example, for this article the Bibclassify module extracted from PSH these terms (followed by the scores reached): *grey literature (40)*, *science (31)*, *www (24)*, *licences (21)*, *plant developmental stages (20)*, *author (19)*, *work (18)*, *data theory (17)*, *documents (15)*, *copyright (12)*, *resources (11)*.

An example of a more advanced automated indexing application is *Maui Indexer*[6]. It also accepts knowledge organisation systems in SKOS format from which the terms are assigned, but unlike BibClassify module it uses machine learning algorithms. Given a sufficiently large set of documents with manually assigned indexing terms from which it can "learn"(i.e. build a model), its indexing performance can be very high[7]. It also provides other features than automatic term assignment (e.g., automatic tagging, topic indexing with Wikipedia).

Both these automatic indexing applications build upon well-structured SKOS format which is essential for their operation because it provides widely accepted formalization of KOS that serve as a standard. It is an example how semantic web technology can improve on traditional ways of KOS distribution. As Vlachidis puts it, *„today available semantic technologies promise to close the gap between formal knowledge structures and textual representations enabling new access methods to information"* (VLACHIDIS, 2009).

2.2 Enhancing user interface

Taking the perspective of the user, knowledge organisation systems enhance the information retriveval system's user interface. The improvements can be made to the search process by search query refinement. The interface may provide the user with suggested spelling corrections or offer auto-completion of entered search terms. The more sophisticated application of KOS is broadening the search using techniques of semantic query expansion.

Knowledge organisation system can be used as a tool for hierarchical navigation and other alternative ways of browsing. Building on the KOS's rich relationships between its terms user can be presented with contextual information, such as related documents. Some of the contemporary digital library software (e.g., JeromeDL[8]) makes it possible for users to assign social tags to documents based on underlying subject organisation. Tags constitute a custom-made access points to the documents in the digital library collection. If KOS is used, tags can be assigned in consistent manner which aligns with user's expectations.

2.3 Linked data

An emerging trend can be observed in the area of knowledge organisation systems as they get involved in building the so-called "semantic web". This leads to the *linked data* paradigm that represents the sum of best practices for exposing, sharing and connecting semantic web data.[9] Linked data is a way of publishing documents and data as mutually interlinked web resources. It utilizes semantic web technologies, mainly Resource Description Framework (RDF)[10] data model and Uniform Resource Identifiers (URIs)[11] as a primary way of resource identification. The way linked data is published is guided by four principles formulated by Tim Berners-Lee (BERNERS-LEE, 2006). They recommend using URIs as resource identifiers, particularly HTTP URIs, so people can look them up using a web browser. Linked data should also provide useful information about described things and links to other resources.

Knowledge organisation systems form an important part of web of linked data. They constitute a layer of interoperability between disparate datasets and can serve as hubs interlinking linked data resources. Rich hyperlinks add to the value of connected resources and increase their accessibility which is particularly important for grey literature documents. Tim Berners-Lee suggests that *„the value of your own information is very much a function of what it links to"* (BERNERS-LEE, 2006). Knowledge organisation systems are also used for annotating resources. This can take the form of semantic tagging, using established formats such as Dublin Core[12], Meaning-of-a-Tag (MOAT)[13] or CommonTag[14]. Tagging is often viewed as putting the web resource into context (WILDE, 2008). Tags join related information and provide links users can follow to broaden the scope of their search for information.

2.4 Simple Knowledge Organisation System

In order to take advantage of the afore-mentioned applications of knowledge organisation systems they should be published as linked data as well. That means converting the KOS to some RDF format. There seems to be a growing consensus in adoption of Simple Knowledge Organisation System (SKOS)[15] which is specifically designed to represent traditional knowledge organisation systems. It provides means for expressing standard relationships that can be found in most thesauri or subject heading systems. SKOS is defined using RDF Schema and Web Ontology Language (OWL) and is expressed in RDF data model. In August 2009 it has gained the status of World Wide Web Consortium's (W3C) working recommendation and it is already widely used in many influential libraries (e.g., the Library of Congress). In June 2009 the list of SKOS implementations was joined by the National

Technical Library (NTL[16]). Its product Polythematic Structured Subject Heading System (PSH) was published in SKOS format under Creative Commons (CC) licencing terms.

3 Polythematic Structured

Subject Heading System

Polythematic Structured Subject Heading System (PSH) is a set of subject headings which is used to describe and search the document by subject. PSH in its latest version 2.1 is bilingual (Czech–English) and the subject headings in both languages are interconnected. It contains over 13500 subject headings and it is divided into 44 thematic sections.

Subject headings are included in a hierarchy of six (or seven) levels according to their semantic content and specificity. Nearly none of mentioned thematic sections do not significantly exceed the others in number of its subject headings. The "depth"of each section is approximately the same.

The whole system is a tree structure and it represents various concepts from the most general to the more specific ones. There are hierarchical, associative ("see also") and equivalence ("see") relations in PSH. The set of subject headings has qualifiers appended. The qualifier has a two-letter-abbreviation form, which represents the belonging to one of the thematic sections.

In comparison with uncontrolled terms and keywords which are created freely, PSH is a controlled vocabulary which is based on binding principles similar to thesaurus construction. Despite this, PSH is a user friendly tool when indexing or searching for documents and information. It stems from the natural language and respects natural word order, and therefore it makes searching and indexing easier. Indeed, *„controlled list of terms ensures consistency in searching and indexing, helping to reduce problems arising from synonym and homograph mismateches"* (TUDHOPE, 2006).

For indexers and searchers, PSH is a listing of terms authorized for use in an indexing system, together with relationships, variants and synonyms, and aids to hierarchical navigation through the subject heading system. It serves as a retrieval tool as well. The bilingualism of PSH guarantee its accessibility on international level. Thanks to the interconnection of each subject heading, PSH can be used to translate terms from Czech to English and vice versa.

Traditionally, the subject indexing of textual documents has been the responsibility of professional librarians and indexers. One of the disadvantages of expert indexing is that it is expensive and time consuming. That is the reason why some university libraries require documents to be indexed by authors (students). In the Czech Republic, subject indexing process of theses and dissertations is presently disunited. It is still mostly an intellectual process, where an author-based indexing is founded on the assumption that author´s interpretation of his/her own work is the authoritative view. To the contrary, the indexer must apply terms which appropriately identify the subject either by extracting words directly from the document or assigning words from a controlled vocabulary.

Among the benefits of using subject heading systems when indexing grey literature (e.g., theses and dissertations) belong the lucidity, easy hierarchical-structure orientation and manipulation, etc. This set of agreed-on terms represents specific concepts, and thereby reduces the number of possible wordings for meaning.

PSH has an intuitive browsing tool which is easy to handle for everyone. By the visualisation of PSH, displaying the hierarchical navigation and relations between terms, could be searching and indexing more interesting and easier. At the same time it is possible to display tags from PSH (metadata snippets - Dublin Core and CommonTag) that can be embedded in an HTML document its semantic description in a machine-readable way. Almost anyone authoring web content (e.g., bloggers) may found a use for it.

PSH can be used as a separate tool which authors refer to in deciding how to tag their documents while indexing. It can be used in authomatic indexing systems to generate a list of suggested terms as tags representing key concepts of documents. There is also a possibility of refining user´s queries for retrieval according to PSH.

The value of PSH in indexing and retrieval systems (particularly in the OPAC and other information and retrieval systems) is well recognized. It can serve for example to correct spelling errors and to provide:

- means by that the use of terms may be standardized,
- controlled hierarchies so that a search can be narrowed or broadened systematically,
- a guide for users of the system for choosing the correct term for a subject search,
- a navigation tool through the set of subject headings and their relations.

PSH is produced by the National Technical Library in Prague (NTL), which guarantee its quality. In the context of the PSH development process the Committee for Coordination of Polythematic Structured Subject Heading System was established. The Committee consists of external and NTL experts and group of representatives from libraries participating in usage of PSH. At the end of the year the Committee members gather for annual meeting to discuss and agree on proposals for improving PSH. Committee members can also suggest new ideas concerning recommendations focused on further development directions and possibilites of PSH. The last committee meeting took place on 10th Decemeber 2009. Among other things, there was a disccussion regarding feasibility of offering PSH under one of the CC licences, which was eventually realized.

4 Creative Commons licences

Creative Commons licences are a product of Creative Commons Initiative, an American non-profit organization founded in 2001. CC licences origin is connected with the problem of protecting but at the same time using work. With the development of the Internet and particulary with the newest communication technologies commonly labeled as Web 2.0, we have been witnessing the rapid growth of "participatory media", whereby both proffesionals and amateur authors create and share digital content with greater ease than before (CHELIOTIS, 2007). CC licences are currently used especially for electronic resources accessible through Internet network. *„The Creative Commons licenses is a perfect example of the sort of copyright changes the modern world needs to come to grips with in the digital age, information should be free to all."* - John Harvey (Flickr photographer) (CASE, 2009b).

Creative Commons' main goal is, in particular, enable legal using and sharing of publicly avaliable author work. In response to a various requests by a broad community of authors CC has introduced a number of licence types to suit different needs (CHELIOTIS, 2007). CC defines the spectrum of possibilities between full copyright and public domain. Authors can choose one of the CC licences thereby provide opportunity to share work with others. If author choose one of those CC licences he will actually give up some rights and allow using his work within the scope of the chosen CC licence. CC licences are not an alternative to copyright, they work alongside it (CREATIVE, 2001). *„It takes the concept of "all rights reserved" and splices it into manageable parts in order to create a "some rights reserved" system."* (FILTER, 2009). However, CC licences must be understood in the context of the existing copyright regime (CHELIOTIS, 2007).

Creative Commons Initiative provides a set of licences composed of requirements and permissions represented by four basic licencing elements - requirement for attribution, sharing derived works under same or similar licence, prohibition of commercial use, and restriction against creating derivative works. In turn for adhering with these requirements users may be permitted to reproduce, distribute, share the licenced work and also make derivative works. Special conditions can be specified for use in developing countries.

The licencing terms are expressed using Creative Commons Rights Expression Language (ccREL)[17] which is RDF-based and can be serialized to fit in various document formats (HTML, PDF). The most common is the use of RDFa (RDF in attributes)[18] form that can be easily included in a web page. Metadata describing CC licencing terms serves as an explicit connection between the document and the abbreviated version of licence. In this way, the document is clearly marked so that users know what they are allowed or restricted to do with it. Formalized metadata also enables indexing by search engines which in turn increases visibility of the licenced work.

Creative Commons presents publishing protocol that can greatly benefit grey literature authors. Their main motivation often is to share and improve access to their knowledge or results of their work, while not being conditioned by the expectation of a financial reward. Instead, they want others to use their work and cite it. *„The attribution requirement in the CC-Licence ensures that access to their work builds upon their reputation"* (HOORN, 2006). CC licences allow authors to skip the intermediaries (e.g., commercial publishers) which is what most grey literature authors want or have to do. This is mostly in accord with related *open access* publication model. *„In fact, a large portion of the available GL [grey literature] is OA [open access]"* (HAWKINS, 2008). For the afore-mentioned purposes, CC licences are an optimal tool for effective dissemination of grey literature.

The first set of CC licences was released in 2002 for free to the public. CC licences have grown at an exponencial rate around the world and expanded over American border. National CC licences versions are based on the generic licence, they are created in national language and adapted to specific system of law.

Czech version of Creative Commons licences was translated by civil association Iuridicum Remedium in cooperation with the Faculty of Law, Charles University and the National Library of the Czech Republic. Czech version of Creative Commons licences was published in April 2009.

Licencing PSH under CC is a result of the Committee for Coordination of PSH meeting. But there were more motivating factors behind the decision to use CC licence for PSH. The National Technical Library as a modern institution recognises that new approches to copyright management should be considered in order to exploit the enormous potential for knowledge distribution proffered by the Internet. The main goal was to make PSH avaliable under an open licence to promote the use and dissemination of PSH.

Key motivating factors to release PSH under CC licence include:

- Encourage library community and also Internet users contribution of knowledge to participate in NTL projects including PSH.
- Dissemination of PSH, where possible, in open formats and as open data.
- Give others easy access and the right to use PSH without seeking NTL´s permission.
- Licencing under CC allowed us to make PSH avaliable to anyone who might be interested in controlled indexing and searching.
- We want to take every opportunity to communicate new ideas broadly and seek feedback because great ideas do not develop in isolation.

Proposal how to improve PSH by publishing it under CC licence was approved immediately after publishing Czech CC licence version. PSH is licenced under the Creative Commons Attribution Non-Commercial Share Alike 3.0 (Czech).[19]

The National Technical Library allows any individual user to copy, distribute, display and perform the PSH. But they must give the original author credit, they may not use this work for commercial purposes and if they alter, transform, or build upon this work, they may distribute the resulting work only under an identical licence that governs PSH.

5 Conclusion

Many target groups may get benefit from PSH.

Our aim is to encourage use of PSH by the wider web audience outside of libraries. We will try to ensure PSH has a role in the present-day ever-changing web environment and will not become an obsolete tool. This involves staying up to speed with current trends and recents developments in knowledge organisation systems area which is most probably shifting towards the semantic web.

To librarians we want to communicate the message PSH is a progressive subject heading system that is there available free to use. In future, our effort is to increase consistency of indexing and searching and improve access to digital collection of grey literature by using PSH, e.g., in the National Repository of Grey Literature in the National Technical Library in Prague.

And because in the future publishing certainly will be no shortage of grey literature we will attempt to convince the ever-increasing number of authors and publishers operating in this field of benefits of using PSH. Knowledge organisation systems such as PSH are able to keep the knowledge these people are trying to share from fragmenting into unmanageable chaos of irretrievable and unaccessible bits and bytes.

Bibliography

- BERNERS-LEE, Tim. *Linked data : design issues* [online]. Published 2006-07-27. Last change 2009-06-18 [cit. 2009-06-10]. Available from WWW: <http://www.w3.org/DesignIssue/LinkedData.html >.
- BINDING, Carl; TUDHOPE, Douglas. KOS at your service : programmatic access to knowledge organisation systems. *Journal of Digital Information* [online]. 2004 [cit. 2009-09-30], vol. 4, no. 4. Available from WWW: <http://journals.tdl.org/jodi/article/view/110/109 >. ISSN 0972-7272.
- *Case study : Blender Foundation* [online]. 2009a. Last change 2009 [cit. 2009-10-02]. Avaliable from WWW: <http://wiki.creativecommons.org/Case_Studies/Blender_Foundation >.
- *Case study : Filter Magazine* [online]. 2009b. Last change 2009 [cit. 2009-10-02]. Avaliable from WWW <http://wiki.creativecommons.org/Case_Studies/Filter_Magazine >.
- *Case study : Monkeyc.net* [online]. 2009c. Last change 2009 [cit. 2009-10-02]. Avaliable from WWW <http://wiki.creativecommons.org/Case_Studies/Monkeyc.net >.
- *Creative Commons* [online]. [2001] [cit. 2009-10-05]. Avaliable from WWW: <http://creativecommons.org/ >.
- CHELIOTIS, Giorgos [et al.]. Taking stock of the Creative Commons experiment monitoring the use of Creative Commons licenses and evaluating its implications for the future of Creative Commons and for copyright law. In *Proceeding of the 35 conference on research communication, Information and Internet Policy (TPRC)*. Virginia : George Mason University School of Law, 2007. Also avaliable from WWW: <http://wiki.creativecommons.org/images/7/71/Taking-stock-of-the-creative-commons-experiment_eng.pdf >.
- CHILDRESS, Eric. Going grey : grey literature and metadata. *Journal of Internet Cataloging*. 2003, vol. 6, no. 3. ISSN 1091-1367 (print). DOI 10.1300/J141v06n03_02.
- DE BLAAIJ, Cees. Grey literature from invisibility to visibility. *Publishing Research Quarterly*. 2004, vol. 20, no. 1, pp. 70–76. ISSN 1053-8801 (print). DOI 10.1007/BF02910862.

- DE BLAAIJ, Cees. Two worlds : about bars and stars in scientific information publishing, an analysis of open source ideology as a means of self-controlled publishing. *Grey Journal*. 2005, vol. 1, no. 1. ISSN 1574-1796 (print).
- HAWKINS, Donald T. Gray literature : what's new on the information landscape. *Information Today*. February 2008, vol. 25, no. 2, pp. 27–29. ISSN 8755-6286.
- HOORN, Esther. *Creative Commons licences for cultural heritage institutions : a Dutch perspective* [online]. Amsterdam : University of Amsterdam. Faculty of Law. Institute for Information Law, 2006 [cit. 2009-09-30]. Available from WWW: <http://www.ivir.nl/creativecommons/CC_for_cultural_heritage_institutions.pdf >.
- KRUK, Sebastian Ryszard; MCDANIEL, Bill (eds.). *Semantic digital libraries*. Berlin, Heidelberg : Springer, 2009. ISBN 978-3-540-85433-3.
- MACKENZIE OWEN, John S. The expanding horizon of Grey Literature. In *Perspectives on the design and transfer of scientific and technical information : proceedings of the 3 international conference on grey literature*. Amsterdam : TransAtlantic, 1998, pp. 9–13. Also available from WWW: <http://cf.hum.uva.nl/bai/home/jmackenzie/pubs/glpaper.htm >.
- *Perspectives on the design and transfer of scientific and technical information : International conference on grey literature, Luxembourg, 13–14 November 1997*. Amsterdam : Transatlantic, 1997. ISBN 90-74854-17-7.
- RABINA, Debbie L. Copyright licenses and legal deposit practices of grey multimedia materials. *Grey Journal*. 2009, vol. 5, no. 1. ISSN 1574-1796 (print).
- SCHÖPFEL, Joachim. Observations on the future of grey literature. *Grey Journal*. 2006, vol. 2, no. 2. ISSN 1574-1796 (print).
- SHIRI, Ali Asghar; REVIE, Crawford. Thesauri on the web : current developments and trends. *Online Information Review*. 2000, vol. 24, no. 4, pp. 273–279. Available from WWW: <http://dlist.sir.arizona.edu/163/01/thesauri.pdf >. ISSN 1468-4527.
- SHIRKY, Clay. *Ontology is overrated : categories, links, and tags* [online]. 2005 [cit. 2009-09-30]. Available from WWW: <http://www.shirky.com/writings/ontology_overrated.html >.
- SIEGEL, Gretta E. Capturing academic grey literature : starting at home. *Publishing Research Quarterly.* 2004, vol. 2, no. 1, s. 62–69. DOI 10.1007/BF02910861. ISSN 1053-8801 (print).
- TUDHOPE, Douglas; KOCH, Traugott; HEERY, Rachel. Terminology services and technology : JISC state of the art review. *UKOLN Newsletter* [online]. 2006 [cit. 2009-09-30], iss. 17. Also available from WWW: <http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf >. ISSN 0963-7354.
- VESELY, Martin [et al.]. CERN Document Server : document management system for grey literature in networked environment. In *international conference on grey literature : grey matters in the world of networked information*. Amsterdam : TextRelease, 2004. GL-conference series, no. 5. Also available from <http://www.greynet.org/images/GL5-Vesely.pdf >. ISBN 90-77484-01-9.
- VLACHIDIS, Andreas [et al.]. *Excavating grey literature : a case study on rich indexing of archaeological documents by the use of natural language processing techniques and knowledge based resources* [online]. 2009 [cit. 2009-09-30]. Preprint for ISKO UK 2009 : Content Architecture: Exploiting and Managing Diverse Resources. Available from WWW: <http://www.iskouk.org/conf2009/papers/vlachidis_ISKOUK2009.pdf >.
- WILDE, Eric; LIU, Yiming. Lightweight Linked Data. In *Proceedings of the 2008 IEEE International Conference on Information Reuse and Integration*. New York (NY) : IEEE Systems, Man, and Cybernetics Society, 2008. Also available from WWW: <http://dret.net/netdret/docs/wilde-iri2008-lightweight-linked-data.pdf >.

References

[1] http://www.textrelease.com/textreleasehome/pressreleases.html
[2] http://sindice.com/
[3] http://psh.ntkcz.cz/skos/PSH6454/html/en
[4] http://nusl.techlib.cz
[5] http://invenio-demo.cern.ch/help/admin/bibclassify-admin-guide
[6] http://code.google.com/p/maui-indexer/
[7] This can be illustrated by an AGROVOC test case – http://code.google.com/p/maui-indexer/wiki/Examples\#Assignment_of_Agrovoc_terms_to_agricultural_documents
[8] http://www.jeromedl.org/
[9] http://linkeddata.org/
[10] http://www.w3.org/TR/rdf-syntax-grammar/
[11] http://www.w3.org/Provider/Style/URI
[12] http://dublincore.org/
[13] http://moat-project.org/
[14] http://www.commontag.org/Home
[15] http://www.w3.org/2004/02/skos/
[16] http://www.techlib.cz/en
[17] http://www.w3.org/Submission/ccREL/
[18] http://www.w3.org/TR/xhtml-rdfa-primer/
[19] http://creativecommons.org/licenses/by-nc-sa/3.0/cz/

https://doi.org/10.26069/greynet-2026-000.559-tgj

Slovak Centre of Scientific and Technical Information

Our mission is to support the development of science, technology and education in the Slovak Republic.

- Operating and improvement of complex information systems for research and development – Central Information Portal for Research, Development and Innovation. (more info: www.vedatechnika.sk)
- Popularization of science and technology in the Slovak Republic – The National Centre for Popularization of Science and Technology in Society. (more info: www.vedatechnika.sk)
- Implementation of the national projects, which are aimed at improving of information security of research and development and services for research and development. (more info: www.cvtisr.sk, http://nispez.cvitsr.sk, http://dc.cvtisr.sk)
- Library and information activities – national information center for support of science and technology and specialized scientific public library of the Slovak Republic. (more info: www.cvtisr.sk)
- Methodical and analytical activities, which are aimed at supporting the management and evaluation of research, development and higher education for the Ministry of Education, Science, Research and Sport of the Slovak Republic.

Contact address:
Slovak Centre of Scientific and Technical Information
Lamačská cesta 8/A, 811 04 Bratislava, Slovak Republic
Tel.: +421 2 69253 102
Fax: +421 2 69253 180
E-mail: cvti@cvtisr.sk
www.cvtisr.sk

https://doi.org/10.26069/greynet-2026-000.559-tgj

Grey Documents in Open Archives *

Daniela Luzi *(Italy)*

Introduction

Science is a social activity [Merton, 1973] based on a cumulative process of knowledge building and sharing. This process relies on the efficacy of an information infrastructure defined as "the technological, social, and political framework that encompasses the people, technology, tools, and services used to facilitate the distributed, collaborative use of content over time and distance" [Borgman, 2007]. The use of ICT (Information and Communication Technologies), and of Internet in particular, is radically changing the way in which scientific research is carried out and consequently the way in which information is shared and exchanged. This has called into question the roles and functions of the principal actors that, in the so called 'print era', contributed to adding value to the process of scientific communication and, at the same time, solicit new petitions for the free circulation of knowledge proposed by the Open Access movement.

Many of these changes are still ongoing. They involve both the use of the new technologies and a cultural shift in the information production and dissemination practices among different scientific communities. A slow and still to be fully negotiated reappraisal of the roles of the different actors in the value chain of scholarly communication is emerging, starting from that of the commercial publishers who propose hybrid models (author pay, institution pay, pay per view) in search of alternatives that do not weaken their income position. Lastly, at the institutional level, both nationally and internationally, a gradual establishment of the principles of free knowledge circulation (starting from those stated in the Berlin Declaration) is underway as it is the proposal of policies, the implementation of which could set the direction of and thus substantially accelerate these changes (take, for example the policies adopted by several national and international organizations to make submission of scientific publications mandatory).

Nevertheless, the conviction is now emerging that ICT is still not being exploited to the full as it has so far merely telematically duplicated the conventional process of scholarly communication; just as the paper-based documents have simply been transformed into analogous digital versions. Projects and research in this sector (Web 2, OAI-ORE) and examples of applications in advanced sectors are now proposing new and promising solutions pointing to possible much more radical changes in the way research is carried out, how information is accessed and then re-used. The challenge lies not so much in rendering all research products accessible but reconstructing the links among the various scientific outputs, thus reproducing the individual phases of the process of scientific enquiry [Van de Sompel et al. 2009].

If these are the premises, outlined here concisely and in a deliberately simplified way, what is the position of Grey Literature (GL) today? There is no doubt, as many have claimed [Gelfand 2005; Banks et al. 2007], that GL is at home in open archives, whether in the form of e-print archives or Institutional Repositories (IRs). Indeed, precisely insofar as scientific GL records and documents the results of the various research phases, its inclusion in the open archives legitimizes it as a scientific artifact and acknowledges its added value. In a scenario of interconnections among the various research products and of the integration between services and information sources, this means that GL has every right to be included in the process of the production and transmission of knowledge worthy of being conserved and disseminated.

GL in scholarly communication

In the print era, GL acted as a channel for the dissemination of scientific information that ran parallel with that of commercial publishing and was not subject to set production and dissemination rules. This lent GL the characteristics of an informal communication which, according to Meadows [Meadows 1998] is "often ephemeral and made available to a restricted audience only". The similarity between this definition and several of the characteristics peculiar to GL, often cited in support of Wood's definition [Wood 1982, Auger, 1993], is quite apparent.

The opposite pole is represented by formal communication, which culminates in a publication "available over long periods of time to an extended audience" [Meadows 1998]. However, there has never been a clear-cut distinction between the two (and the development of GL is evidence of this) and it is currently dwindling even further in the

* First published in Grey Literature in Library and Information Studies. – Berlin : De Gruyter Saur, 2010.

context of digital media and networked communication. For this reason, in the wider acceptance of the term 'scholarly communication' [Borgman, 2007] defined as the study of "formal and informal activities associated with the use and dissemination of information through public and private channels", GL can provide a good vantage point from which to observe the continuum – whether pre-publication [Harnad 1990] or electronic publication [Kling, 1999, Borgman 2000] – of information exchange activities underlying scientific inquiry.

As part of the renewed interest shown in the ongoing changes in scholarly communication, several recent studies [Pepe et al, 2009] made use of the technique of descriptive laboratory accounts of science–in-action in order to analyse the various artifacts produced by specific scientific communities in the different phases of the life cycle of research activities. These studies have confirmed that in each phase numerous artifacts are produced, each with its own role, specific content and a different communicative function. They also highlighted their continual, progressive updating and enhancement by which they are transformed into new products, ultimately arriving at the "formal" artifact that can be condensed into a publication.

Van de Sompel [Van de Sompel 2004] describes these artifacts as units of scholarly communication that "reflect the changing nature of the information assets produced and consumed in the scholarly endeavors". Borgman [Borgman, 2007] describes the continuity of scholarly communication, pointing out the gradual shaping and reshaping of information intended for specific recipients. Taking these considerations as a starting point, GL can be described on the basis of the functions it performs within scholarly communication. Thus – despite a degree of overlap and hazy outlines – report literature accomplishes the function of describing detailed results of specific research phases, while conference papers and preprints (generally *reshaped* from report literature) are focused on communicating them to a specific audience. In the context of higher education, theses perform the function of certifying the acquisition of an academic qualification but at the same time report the results of the experimental research undertaken by the candidate or else describe the state of the art of a specific topic, whereas courseware materials, which are also developed as a function of the academic level of the reference audience, provide a systematic framework for the knowledge so far gained in a given disciplinary field.

Side by side with these artifacts, modes of web-based interaction have grown up that increasingly resemble face-to-face communication and the collaborative construction of knowledge bases (the well known and extensively consulted wikis). Starting from the first bulletin boards, nowadays several blogs (for example those following the Open Access debate) embody one of the many tools used by virtual communities of scholars to share open peer commentary and documents. From the point of view of circulation, blogs sometimes become actual sources of information that are used to keep up to date on the developments in certain topics and to follow the progress of the ongoing discussion. These tools open up new horizons for scholarly communication (and also for other communications) and give form to constantly growing and transformed grey information or grey content, which deserves a more detailed and specific analysis.

From report literature to scientific artifacts

GL develops in practically all disciplines in response to a need for scientific communication, ranging from that of providing a detailed documentation of research results, without the space limitations of journal articles, to that of reducing the time between information diffusion and their actual publication. These needs have formed the basis of many of the initiatives aimed at setting up open access archives that have often viewed GL documents as a test bench for trying out new dissemination methods and have generally involved the same actors as those who contributed to the development of GL.

The informal nature of GL, previously considered to be one of its main limitations, has actually allowed for both a transformation and a constant increase in the number of document types which corresponded to specific information needs depending on the disciplinary context involved. As mentioned above, each document type usually contains specific information (although variations in the GL field area are indeed the rule). Report literature represents the basic nucleus of paper-based GL, the various types of technical papers (interim reports, research memoranda, working papers) generally describe a particular phase of the research activity and report the results, providing a detailed documentation of data and processing, as well as of the procedures and methods used to analyse them. The absence of publishing rules restricting the length of the articles makes it possible to give a detailed description of the data and procedures adopted. These documents often represent the only information source in which it is possible to find these data that are necessary to duplicate and verify the research objectives and results, and possibly draw new insights from them.

What in the print era made up the content of technical reports or the annexes thereto is

now accessible in the form of "compound units" [Van de Sompel et al. 2007], which is also denoted as "datument, a compound document where all the compounds (data, text, software images, links) are part of the whole" [Murray-Rust 2008]. The transposition of these documents into a digital environment allows their information content to be enhanced considerably. It offers the opportunity of including in the electronic document hyperlinks not only to other bibliographic sources, but also to other types of scientific artifacts, such as simulations, videos, data sets, original lab notebooks and even software used to display and/or further process such data. In this way both research results and the process by which they have been obtained are *reproduced*.

In data-intensive and highly collaborative disciplinary fields (such as molecular biology, earth and space sciences, but also in the social sciences) access to these data become essential. This certainly raises questions related to copyright, authorship and access licenses (and the OA movement has in recent years included in its own agenda also the issues of free access and re-use of datasets), as well as the need to tackle problems linked to the storage and retrieval of this information.

If the digital document is transformed into a compound document, it is necessary not only to find suitable techniques of dataset retrieval from which one or more documents may be prepared, but also link together the various different document types, ranging from technical reports, degree theses, and papers delivered at congresses, down to articles published in commercial journals, each of which accessible from different network locations. The link between the various documents and the data on which they are based means that each scientific artifact serves as an entry point to the set of related artifacts [Pepe et al. 2009]. This is useful for reconstructing the entire research process as well ad capturing the progressive shaping and reshaping of scholarly communication.

Initiatives in this sector are found both at the theoretical level and in systems developed in specific disciplinary fields. The former are being developed by the Open Archives Initiative – Object Reuse and Exchange protocol [OAI-ORE 2008] which published the specifications for handling aggregations of compound information objects of web resources in 2008.

Several systems, which would warrant a separate treatment, point to a growing number of web sites run by research organizations but also by commercial publishers. Here only a few examples will be given. In the field of biomedical and life sciences a number of examples exist, including Nature Proceedings [Nature Precedings], a free service produced by the publishers of *Nature,* which collects pre-publication research and preliminary findings in the biomedical and life sciences. Also within life sciences, BioLit [BioLit Project; Fink et al 2007], set up by the University of California, supplements the articles published by the Public Library of Science (PLoS) with the information contained in the Protein Data Bank (PDB), while another archive, SciVee allows the open uploading of published articles accompanied by the relative video or podcast presentations. In astrophysics mention must be made of the NASA funded Smithsonian Astrophysics Data Systems (ADS) which collects both astronomic and physics literature and link it to data collected by space missions and ground-based observations [SAO/NASA ADS; Eichhorn et al. 2006]. Lastly, also in the social sciences, the Council of European Social Science makes the CESSDA archive available, which allows the retrieval of the datasets and variables of sociological surveys, longitudinal studies, census data collected in the various European countries.

From the preprint culture to Open Access

Preprint culture has generally been attributed to specific scientific communities, in the first instance of physicists and computer scientists. In actual fact, many studies, beginning with those of Gavey and Griffith [Garvey et al. 1967] in the field of psychology, indicate that above all in those sectors in which no short-term commercial applications are to be expected, there is a widespread attitude among scholars to consult and exchange preprints or other types of document that have not yet been formally published in journal articles, in order to keep abreast of the latest research developments and to seek comments via private circulation to friendly reviewers.

The physics sector may be considered emblematic for various reasons. It must not indeed be overlooked that physicists were among the first users of networks even before the advent of the Internet. The World Wide Web was conceived at the European Organization for Nuclear Research (CERN), and the first US web site was opened a few months later (December 1991) at the Stanford Linear Accelerator Center (SLAC) precisely to gain remote access to the SPIRES-HEP (Stanford Public Information Retrieval System-High Energy Physics) database, one of the richest electronic archives set up to handle preprints.

The SPIRES database is a good example of the tight collaboration network that exists among libraries, as well as the productive interaction between libraries and their own local and remote users. This database was developed to facilitate the distribution of preprint lists diffused by post by the SLAC library in the early 1970s and was progressively enhanced thanks to the contributions of both libraries of similar institutions and SLAC researchers,

who developed advanced tools to support the timely diffusion of GL documents [Carroll et al. 1994; Kreitz et al. 1996, O'Connel 2000]. These collaborations and synergisms were able to nourish the preprint culture in that they triggered a virtuous circle between the demand for timely information and a wide range of information sources on offer, which simply reinforces the attitudes of sharing and exchanging information strongly.

This is the environment in which the well-known ArXiv archive developed. In 1991 Ginsparg created a centralized system for the electronic distribution of e-prints at Los Alamos National Laboratory. Designed for a group of about 160 High Energy Physics researchers, the future ArXiv e-print archive expanded rapidly. In just a few months it was not only extended to 1000 users (preprint readers and submitters) and adopted in other areas of physics, but was also introduced and used in other disciplines (mathematics, computer science, linguistics and cognitive sciences, and even in economics).

Initially the system was conceived as a central preprint archive, that is, the type of GL document closest to the final versions of a journal article, not yet subject to copyright and not yet subjected to the formal peer-review process. E-prints apparently mark the transition from hard-copy preprint to the electronic preprint. In actual fact the ArXiv archives, and all those based on this model, soon turned into archives in which GL and conventional literature overlapped. For example, in the current version of ArXiv, full-text access of the document is necessary to verify whether, together with the univocal number attributed to the e-print, there is also an indication of the periodical in which it was published. The e-print definition given by the Joint Information Systems Committee (JISC) [Swan et al. 2005] on the one hand highlights its function ("a digital duplicate of an academic research paper that is made available on line as a way of improving access to the paper") and on the other, emphasizes peer review, which becomes the only, albeit important, difference between GL and conventional literature. Currently an un-referred preprint is distinct from a peer-reviewed postprint. Many publishers nowadays also allow authors to self-archive postprints on their own web page or in open archives, considering them *different* from the publisher-generated format.

The main novelty lies in the fact that the ArXiv archive created a self-sufficient communication model, without intermediaries, which enhances and exploits the interactive aspect. So much so that it seems that the phases of acquisition, storage and dissemination coincide and the roles of author and reader, and even of metadata supplier, now overlap.

The self-consistency of this communication model is highlighted also by Van de Sompel [Van de Sompel et al. 2004], who analyses it using the value chain functions as applied by Roosendaal and Guerts [1997] to scholarly communication. Self-archiving corresponds to the *registration* phase which allows claims of precedence for scholarly funding, the *certification* function, conventionally performed by peer review, is carried out according to Van de Sompel through the procedures of endorsement of potential submitters by peers, but probably also by the reputation of the institutions and by that of its scholars in a scientific community accustomed to collaborating in experiments involving a large number of researchers. The *awareness* function is performed by the dissemination of scientific content that is freely accessible online, by alerting services and by allowing search engines to index content, while the *archiving* function is "based on ensuring adequate redundancy through the operation of a network of separately controlled systems". However, *Rethinking scholarly communication* does not mean giving priority to the ArXiv models over the conventional process of scholarly publishing, but once again designing a flexible and interconnected research infrastructure that can be enhanced by the contributions received from different services and actors, capable of reflecting the information needs of the various scientific communities in order to promote the advancement of scientific knowledge.

The prominence to which the Ginsparg system rapidly rose was not dependent solely on the number of e-prints submitted or the number of archive accesses, but was the result of the convergence of many factors. The ArXiv actually developed at a time in which library budget cuts and the soaring subscription costs (the socalled 'journal crisis' and 'permission crisis') revealed the critical weaknesses of the conventional model of scholarly publishing. This came about at the same time as the expansion of Internet, which instead opened up the possibility of timely information dissemination to a potentially unlimited number of users at relatively low cost.

The best practice represented by the ArXiv archive was therefore a point of reference in the development of the Open Access movement. Converging on it were both the needs of free access and timely dissemination of information pursued by the scientific community and those of libraries, whose role of selection, acquisition and conservation appeared to be strongly limited by the acquisition of bundled packages of journals. Since the Santa Fe convention [Van de Sompel et al 2000], the OA movement took up a proactive stance, identifying organizational structures (service and data providers) and technical instruments (from open source software to the development of the specifications of the Protocol for

Metadata harvesting (PMH), enabling archive interoperability) in order to propose an "open scholarly publication framework on which both free and commercial layers can be established". By successively promoting a Green Road (self-publishing by depositing articles in open archives) and a Golden Road (creation of open access journals), the OAI outlined a process of development of scholarly communication in which scholars and libraries repossessed a part of the research products, thus becoming the direct managers thereof. In this way they contribute to counterbalancing a market dominated by the hegemony of a small number of publishers.

The ArXiv currently contains more than 550,000 e-prints, with an annual increase of about 55,000 documents and continues to be a system that is "scientist driven: articles are deposited by researchers when they choose – either prior to, simultaneous with, or post peer review" [Ginsparg 2007]. About 90% of High Energy Physics preprints are immediately and freely accessible online [SCOAP 2007].

The large number of archives based on the Ginsparg model (including the historical RePec in economics, Cogprint for the neurosciences and E-LIS in information science), although not attaining the same percentages as the HEP preprints, continues to display a constant increase in the number of documents. So much so that many [Borgman 2007; Swan 2005] have pointed out that authors showed a greater propensity to submit their work to these thematic archives rather than self-archiving their works in Institutional Repositories (IRs). The impact of these archives is measured in terms of access statistics, document downloads, and by the list of most cited preprints. A growing number of studies [Harnad et al 2004, Soong 2009] also indicates that open access papers are read and cited more frequently as they are freely and more rapidly accessible. Lastly, the preprint archives are provided with a whole series of gateways that facilitate research in a number of preprint archives (for example, E-Print Network set up by the US Department of Energy) or autonomous systems of citation indexing, such as CiteSeer. These initiatives help indicate alternatives to the conventional process of publication, allowing access at many more points and providing parallel services that enhance the scholarly value chain.

From institutional repositories to the interconnected knowledge network

While the creation of e-print archives may be considered as a bottom-up initiative, managed directly by a specific scientific community, that of the Institutional Repositories (IRs) marks the official commitment of the universities and research institutions to making their own scientific artifacts freely available. The current tendency is to classify open archives as disciplinary or thematic (the original e-print archives) and the repositories as institutional /departmental, governmental or aggregating IRs [OpenDOAR]. The distinction is made herein for chronological purposes, as IRs derive from the former and above all because from the standpoint of GL, IRs provide a *natural home* for GL. This is due to a series of factors.

The commitment of the institutions has both a political and an operational value. The first aspect is apparent in the commitment to the OA movement as formally expressed the research institutions' endorsement of the Berlin Declaration. This commitment expresses the institutions' intention to regain a proactive role in scholarly communication. Lynch actually views IRs "as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication" [Lynch 2003]. A similar, and even more radical, stance is taken by Crow [Crow 200] in the oft-cited SPARC (Scholarly Publishing and Academic Resources) position paper stressing the role of IRs as an instrument that "increases competition and reduces the monopoly power of journals, and brings economic relief and heightened relevance to the institutions and libraries that support them". In this way, IRs become also "a digital version of the traditional university press" [Swan et al 2005], the beneficiaries of which, in addition to the scientific community and of course the public at large, are also the libraries.

The aim of the IRs is to develop "a set of services that the university offers to the members of its community for the management and dissemination of digital materials created by the institutions and its community members" [Lynch 2003]. IRs thus represent a comprehensive showcase of the scientific, teaching and cultural activities of a scientific institution. Therefore, the digital materials that are to be rendered freely available include all types of research products (preprints, postprints, theses, conference papers, monographs, research data sets and databases), teaching materials (courseware, lecture notes, etc.) and of course new kinds of grey contents. Each scientific community can decide what kind of collections are to be self-archived, the relative format (whether full-text or bibliographic references) and can ultimately indicate the rules governing access (for instance, by restricting full-text accessibility exclusively to members of the institution).

The organization of scientific content into collections ensures greater visibility of GL, places it in the context of the other artifacts produced by a given community and allows each GL document type to be linked to the appropriate metadata that will facilitate access to and

cross-searching among the various repositories and/or search engines. The latest survey of the EU sponsored Digital Repository Infrastructure Vision for European research Project (DRIVER) [Vernooy-Gerritsen et. al 2009] reports that 62% of European IRs consist of GL, 39% of which are theses, 14% proceedings and 9% working papers, compared with 34% journal articles and 4% books and book chapters. These figures aroused considerable interest in that the survey claimed that one of the functions of IRs is to become "a source for grey literature" for users, as well as an "alternative route to toll-access literature" [Vernooy-Gerritsen et. al 2009]. Furthermore, also the figures given in the Directory of Open Access Repository [openDOAR], indicate that of the 1532 IRs surveyed, 50% contain theses and dissertations, 41% unpublished reports and working papers, 15% learning objects, and 4% data sets.

The organizational model put in place by IRs envisages a more central role for libraries, compared with the one played in the original e-print archives. Indeed the library is called upon to participate both in the repository design phase, in which librarianship contributes to defining the collections and to identifying the metadata required to describe them. This is the case both at the stage in which the repository becomes operational and library expertise is needed to validate the data and to identify the long-term strategies for the preservation of digital formats and/or for any digitization of previous collections of paper-based documents. The library is also assigned the tasks of assisting researchers in self-archiving operations and of providing support for the activities of open access advocacy. This is a role that entails ever closer collaboration between the libraries and the reference scientific community, at the same time projecting the latter towards a larger network environment.

No direct link has ever been found between IR development and the previous experience of libraries in handling catalogues and archives explicitly dedicated to the collection of GL [Di Cesare 2006] or their participation in international initiatives aimed at GL dissemination (a typical example is the SIGLE database). There is no doubt that IRs benefit from the previous development of the digital libraries or OPAC and from the libraries' entire accumulated experience of scientific documentation management and thus it can be inferred that they also benefit from the previous complex management of GL. It would be interesting to investigate whether a *GL management culture* coexisted with the preprint culture. This sometimes emerges in the feasibility studies regarding IRs [Lambert 2006] or in setting up IRs based on existing GL collections [Anderson et al. 2007]. In the current debate on OA, increasing mention is made of GL and in particular of the new forms of GL or grey contents. As well as increasing the visibility of GL, this fully justifies its inclusion in research infrastructure networks.

Concluding remarks

During the print era, GL provided an informal channel for the dissemination of scientific information thanks to the development of its own acquisition and distribution network, albeit for a limited number of experts. Research institutions that were particularly sensitive to its information value set up specialized services for its collection and dissemination; researchers used their own transmission channels and/or were helped by specialized librarians to 'dig it out'. The latter undertook to identify the most suitable ways and means to index and catalogue it. Lastly, several international initiatives, notably SIGLE, funded programs aimed at supporting cooperation among the European countries in the collection and dissemination of GL.

Nowadays open archives are one of the new channels of GL dissemination that enormously amplify its user basin and situate it deservedly in the continuum of scholarly communication. The coexistence of GL with conventional literature actually provides an *ideal, complete coverage* of the research results of any given scientific institution or disciplinary community. It also enables it to be included in the wider debate on Open Access and in the numerous initiatives being developed around this movement.

The distinction between GL and conventional literature is becoming increasingly blurred: GL, in its various forms, is "made public" [Borgman 2007] on the web at a previously unimaginable speed, and the main difference vis-à-vis conventional literature is inherent mainly in the fact that GL is not subjected to any formal peer-review process. New forms of grey contents, which represent essential sources of information for the advancement of knowledge, are opening up new fields of study concerning the conduct of scientific research, as well as the need to preserve and disseminate these artifacts. In this framework, the experience acquired by GL specialists can make a substantial contribution to facilitating and orientating the constant evolution of scholarly communication.

References

Anderson N., G. Hodge, A. Japzon (2007), Harnessing NASA Goddard's Grey Literature: the Power of a Repository Framework. - In: Publish Research Quarterly 23 pp. 58-64.

Auger, C. P. (1993), Information Sources in Grey Literature, 3rd ed. London: Bowkwe-Saur.

Banks, M. A., C. de Blaaij (2007), Implications of copyright evolution for the future of scholarly communication and grey literature. In: Eighth International Conference on Grey Literature : Harnessing the Power of Grey, 4-5 December 2006 / compiled by D. J. Farace / compiled by D. J. Farace and J Frantzen. – Amsterdam: TextRelease, Feb. 2007, pp 43-52.

Berlin Declaration (2003), The original Berlin Declaration is located at <http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>.

Borgman C. L. (2000), Digital Libraries and the Continuum of Scholarly Communication. – In: Journal of Documentation 56 (4) July 2000, pp- 412-430.

Borgman, C. L. (2007), Scholarship in the digital age: information, infrastructure, and the Internet. Cambridge: the MIT Press.

Carroll B. C., G. A. Cotter (1994), A new generation of grey literature: the impact of advanced information technologies. - In: First International Conference on Grey Literature: Weinberg Report 2000, 13-15 December 1993 / compiled by D. J. Farace. – Amsterdam: Amsterdam: TransAtlantic, 1994, pp 5-17.

Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper, Washington: SPARC, 37 p. http://www.arl.org/sparc/publications/papers/index.html.

Di Cesare, R. D. Luzi, R. Ruggieri (2006), Open Archives and SIGLE participation in Italy: Is there a subtle link between the two experiences? - In: An International Journal on Grey Literature 2 (1) spring 2006.

Eichhorn, G., A. Accomazzi, C. S. Grant, E. A. Henneken, D. M. Thompson, M. J. Kurtz, S. S. Murray (2006), Connectivity in the Astronomy Digital Library, ArXiv preprint. arXiv:cs/0610008. http://arxiv.org/abs/cs/0610008. To appear in Library and Information Systems in Astronomy V.

Fink, L. L., P. E. Bourne, (2007), Reinventing Scholarly Communication for the Electronic Age. - In: CTWatch Quarterly 3 (3) Aug 2007, pp. 26-31. http://www.ctwatch.org/quarterly/.

Garvey, W. D., B. C. Griffith (1967), Scientific Communication as a social system. – In: Science 146 (3652) 1967, pp. 1011-1016.

Gelfand, J. (2005), "Knock, Knock:" Are Institutional Repositories a Home for Grey Literature? - In: Sixth International Conference on Grey Literature: Work on Grey in Progress, 6-7 December 2004 / D. J. Farace and J Frantzen. – Amsterdam: TextRelease, Jan. 2005, pp10-15.

Ginsparg, P. (2007), Next-Generation Implications of Open Access. CTWatch Quarterly 3 (3) Aug 2007, pp. 11-18. http://www.ctwatch.org/quarterly/.

Harnad S. (1990), Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. - In: Psychological Science 1, pp 342- 342, http://cogprints.org/1581/0/harnad90.skywriting.html.

Harnad, S. T. Brody, (2004), Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. - In: D-Lib Magazine 10 (6) June 2004. http://www.dlib.org/dlib/june04/harnad/06harnad.html.

Kreitz, P. A., L. Addis, H. Galic, T. Johnson (1997), Collaborative International Control of High-Energy Physics Pre-Prints. - In: Second International Conference on Grey Literature: Grey Exploitations in the 21st Century, 2-3 November 1995 compiled by D. J. Farace. – Amsterdam: Amsterdam: TransAtlantic, 1996, pp. 33-41.

Kling, R., G. McKim (1999), Scholarly Communication and the Continuum of Electronic Publishing. - In: Journal of the American Society for Information Science, 50 (10) 1999, pp. 890 – 906.

Lambert S., B. Matthews, C: Jones (2006, Grey literature, institutional repositories and the organizational context. – In: Seventh International Conference on Grey Literature : Open Access to Grey Resources, 5-6 December 2005 compiled by D. J. Farace and J Frantzen. – Amsterdam: TextRelease.

Luzi, D. (1996), Internet as a new distribution channel of scientific Grey Literature. The case of Italian WWW servers. - In: Second International Conference on Grey Literature: Grey Exploitations in the 21st Century, 2-3 November 1995 compiled by D. J. Farace. – Amsterdam: Amsterdam: TransAtlantic, 1997, pp 42-55.

Lynch C. A. (2003), Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthly Report 226 Feb 2003.

Meadows, A. J. (1998), Communication Research, San Diego, CA: Academic Press.

Merton, R. K. (1973), The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press.

Murray-Rust, P. (2008), Open Data in Science. – Nature Precedings http://precedings.nature.com/documents/1526/version/1. Published in Serials Review, March 2008.

O'Connel H. B. (2000), Physicists thriving with Paperless Publishing. ArXiv preprint. ArXiv:physics/0007040v2. http://arxiv.org/abs/physics/0007040. Invited talk at the AAAS Meeting, February 2000 in Washington, DC.

Open Archives Initatives – Object Reuse and Exchange protocol (OAI-ORE), (2008), Ore Specification and User Guide, http://www.openarchives.org/ore/1.0/toc.

Pepe A., M. Mayernik, C. L. Borgman, H. Van de Sompel (2009), From Artefacts to Aggregations: Modelling Scientific Life Cycles on the Semantic Web. ArXiv preprint. arXiv:0906.2549v3, To appear in the Journal of the American Society for Information Science and Technology (JASIST).

Roosendaal, H. P. Geurts (1997), Forces and functions in scientific communication: an analysis of their interplay. - In: CRISP 97, Cooperative Research Information Systems in Physics, http://www.physik.uni;oldenburg.de/conferences/crisp97/roosendaal.html

SCOAP Working Party (2007). Toward Open Access Publishing in High Energy Physics. Executive Summary of the Report of the SCOAP3 Working Party, CERN 2007. http//scoap3.org/files/Scoap3WPReport.pdf.

Smithsonian Astrophisics Data Systems (SAO/NASA ADS), http://adswww.harvard.edu/.

Soong, S. S. Measuring Citation Advantages of Open Accessibility. D-Lib Magazine 15 (11/12) Nov. 2009. http://www.dlib.org/dlib/november09/soong/11soong.html.

Suber, P (2009), News from the Open Access. http://www.earlham.edu/~peters/fos/fosblog.html

Swan A., P. Needham, S. Probets, A. Muir, A. O'Brien, C. Oppenhaim, R. Hardy, F. Rowland, (2005a), Delivery, management and access model for E-prints and Open Access Journals within further and higher education. A Joint report by The Electronic Publishing Innovation Centre in partnership with Key Perspective Limited. Loughborough: EPIC, 121 p. http://eprints.ecs.soton.ac.uk/11001/.

Swan A., P. Needham, S. Probets, A. Muir, C. Oppenhaim, A. O'Brien, R. Hardy, F. Rowland, s. Brown (2005b), Developing a model for e.-print and open access journal content in UK further and higher education. – In: Learned Publishing 18 (1) Jan 2005, pp. 25-40.

Vernooy-Gerritsen, M., G. Pronk, M. van der Graaf (2009), Three Perspectives on the Evolving Infrastructure of Institutional Research Repository in Europe. - In: Adiadne Magazine 59 Apr. 2009 http://www.ariadne.ac.uk/issue59/vernooy-gerritsen-et-al/.

Van de Sompel H., C. Lagoze (2000), The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine 6 (2) Feb. 2000.

Van de Sompel, H., S. Payette, J. Erickson, C, Lagoze, S. Warner (2004), Rethinking Scholarly Communication: Building the System that Scholars Deserve. - In: D-Lib Magazine, 10 (9) Sep 2004. http://dx.doi.org/10.1045/september2004-vandesompel.

Van de Sompel, H., C. Lagoze C. (2007), Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. - In: CTWatch Quarterly 3 (3) Aug 2007, pp. 32-41. http://www.ctwatch.org/quarterly/.

Van de Sompel, H., S. Warner, C. Lagoze, R. Sanderson, M. L. Nelson, P. Johnston (2009), Adding eScience Assests to Data Web. - In: Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009). ArXiv preprint. arXiv:0906.2135v1; http://arxiv.org/abs/0906.2135.

Wood, D. (1982), Grey literature – the role of the British Lending Division. - In: ASLIB proceedings, 32 (11/12) 1982, pp. 459-465.

Cited web sources

BioLit Project, http://biolit.ucsd.edu/doc/.
CiteseerX - http://citeseerx.ist.psu.edu/.
Council of European Social Science (CESSDA) http://www.cessda.org/.
E-print Network - http://www.osti.gov/eprints/
Nature preceding, http://precedings.nature.com/.
Open Archives Initiative, http://www.openarchives.org.
OpenDOAR - http://www.opendoar.org/.
SAO/NASA ADS - http://adswww.harvard.edu/

https://doi.org/10.26069/greynet-2026-000.559-tgj

Grey Literature and Computational Linguistics: From Paper to Net *

Claudia Marzi, Gabriella Pardelli, and Manuela Sassi *(Italy)*

Abstract

The advent and exponential development of the World Wide Web has led to an increasing availability of unstructured knowledge and distributed information sources, meeting general public requirements that are hardly addressed by other more traditional information channels. This trend has concurrently raised a considerable interest in the application of Computational Linguistics (CL) methodologies to document access and retrieval, as they offer the unprecedented opportunity to make the subjective, user-centred information demands of Net citizens meet the ever changing and heterogeneous information flow of the web. Over the last five years, more and more Italian Universities have introduced CL courses into their Humanities curricula, making available on-line teaching materials, tutorials and language engineering software that appear to supply the lack of offer from traditional Italian publishing houses. In this paper, we consider in some detail the role played by this type of Grey Literature in bringing up a wider and increasingly more aware community of web users in Italy.

1. Introduction

1.1 Computational Linguistics and Language understanding

Computational Linguistics (CL) and Natural Language Processing (NLP) have profoundly changed the way we look at human language as a subject of scientific inquiry, shifting emphasis from abstract knowledge to real usage (Manning & Schütze 1999).

Understanding language requires the ability to master a heterogeneous system of manifold skills, based on the processing of complex information structures in context (Bybee & Hopper 2001, Jackendoff 2002). In turn, these structures may vary, depending on the speaker's communicative intentions and purposes (Barsalou 1999). In this specific sense, knowledge of language can no longer be decoupled from "doing things with words" (e.g. reading, learning, recalling, guessing, judging etc.). The application of computer technologies to issues of language understanding and production epitomizes such a profound change of perspective in the most exemplary way. CL concerns itself with the empirical testing of language models and has contributed, over the last 15 years, to shedding considerable light on the interplay between linguistic abilities and general cognitive functions such as inference, classification and learning (Mitchell 1993, Guarino 1998). Awareness of these issues can have a tremendous impact on the general-public daily demands for real-time, goal-oriented and personalized information paving the way to a new generation of web users.

To date, the vast majority of information available on the web is conveyed through a flood of largely unstructured and uninterpreted text material, ranging from large digital archives to blogs, electronic newswires and dedicated web pages. Flexible, intelligent access to such a digital haystack is a logical precondition to its very existence: information needles are non existing if the haystack is not structured, disseminated and made available to respond, in real time, to the personalized needs, queries and ever changing goals of daily web users (Lawrence & Lee 1999).

1.2 Language, ontology and web contents

Gaining intelligent access to web-based contents presupposes, among other things i) a formal representation (ontology) of the knowledge areas of interest, ii) an algorithm for classifying a new text item according to known ontological categories, iii) an automated procedure to relate terms and complex relations among terms to knowledge categories, iv) ways for automatically indexing contents on the basis of the terms and relations they contain, v) ways for querying a document repository by terms, relations and concepts (Lancaster 2003, Buitelaar, Cimiano & Magnini 2005).

Clearly, non professional web users can profit from ontology-based and language-based technologies for document access and management. Although considerable progress has been made in this area, however, the search for relevant web-based information is still either a) channelled through the strictures of largely pre-defined, context-free, fully-

* First published in the GL11 Conference Proceedings, March 2010

interpreted document ontologies, or b) limited to simple text queries, containing key words and relatively unstructured word patterns connected through few logical operators. Both modes of access have serious limitations.
Document ontologies are formal representation tools for sharing, categorizing, normalizing and indexing knowledge. They are typically highly structured, explicit and symbolic, but do not necessarily reflect the personal viewpoint of a web user, which is often driven by goal-oriented needs and may be unaware of encyclopaedic knowledge.
On the other hand, projecting text queries onto the web is often based on pattern-matching techniques, searching for surface text strings in a few limited variants. This is very efficient and effective for many purposes, but cannot go beyond surface text information. For example, pattern-matching cannot distinguish between the different uses of the English homonym “bank”.
More advanced algorithms for intelligent document access are currently narrowing down the gap between string-based and concept-based queries of web-based content. These algorithms call for a tighter integration between ontological and linguistic knowledge: language interpretation certainly presupposes knowledge about the world, but much world knowledge is acquired through language.
Since language and ontological knowledge are two complementary ways of representing information, future generation search engines will face the challenge of dealing with flexible and open information, which is structured but goal-directed, explicit but personalized. In this scenario, web users will rely more and more often on language queries to look for information, but they will be supported by more structured and shared ways for content indexing and access.

There is another deep sense in which current progress in language technologies can help to narrow down the persisting gap between the demand for personalized information needles and the unlimited offer of unstructured haystacks on the web. Awareness of issues of automated text interpretation and document understanding may represent an altogether different way of narrowing this gap. Web users that understand NLP algorithms are certainly in a good position to use this understanding for better queries. For example, it is often possible to avoid unwanted ambiguities in a query by simply providing it with more context. Adding the word ‘money’ to a query containing ‘bank’ is useful to convey the intended sense of financial institution. In other cases, the use of synonyms in the query can broaden up the range of retrieved documents to a significant extent. We contend that grey literature can play an important role in increasing the general public awareness of these issues. In the remainder of this paper we investigate the potential impact of web-based dissemination of Italian CL literature onto a new generation of non professional web users.

2. Teaching Computational Linguistics in Italy

2.1 CL and the Web

The increasing popularity of Computational Linguistics, over the last ten years, as a knowledge area of general interest has gone hand in hand with the exponentially growing need to access web-based information by daily web users. This is also witnessed by the number of courses on CL that are nowadays active in the curricula of most Italian universities, and also meets a non specialist demand for a more aware use of language as a vehicle for querying and accessing web-based information.
Such an upsurge of interest, however, has not been accompanied by a comparable growth in the number of Italian books devoted to the topic, as traditional publishing houses are typically conservative and rather reluctant to explore new scientific domains and nonetheless so far capable of addressing the general public need for a more aware use of current technologies only to a very limited extent.
This fact, together with the persisting difficulty to read and understand the English technical literature by Italian undergraduate students, has created an acute demand for accessible Italian grey literature on the topic. To date, the amount of unpublished recommended reading material for taking CL university exams largely exceeds the white literature available on NLP issues.

2.2 Grey CL literature

To make it up for comparative shortage of white CL literature in Italian, CL courses have sprouted dedicated web sites providing tutorials, exercises, power-point presentations, language engineering software and other teaching materials. The increasing availability of on-line materials on language technologies provides also a potential access to a wider community of users, by offering introductory information for a better understanding of i) aspects of computer architecture and functioning, ii) issues of digital text encoding and document representation, iii) aspects of text browsing with personalized search patterns,

iv) issues of document mark-up and classification, v) fundamentals of document content indexing, vi) introductory notions on database management systems.
The trend is setting up high standards for web-based publishing and provide as well an important meeting point between academic information providers and non-academic information consumers, by modifying the general public attitude towards computer-based information access and by prompting more personalized ways of accessing web-based information; as a more intelligent and personalized use of the most popular search-engines ties in with a deeper understanding of how texts are automatically searched and indexed and how text information is internally represented and eventually made available by computers.
A new generation of self-taught web users is getting more and more aware of the potentials and limits of the current language-based technology, in these novel scenarios for information access based on integration of language-specific abilities and language-specific cognitive functions.
When people understand better and better what sort of queries make sense in what contexts and why, they will most likely use this knowledge to avoid the strictures of precompiled information repositories and ontologies and to exploit the web in more personalized and creative ways.

3. Case study

3.1 “Informatica Umanistica” university course in Pavia

The university course “Informatica Umanistica” at the Faculty of Humanities in Pavia for the academic year 2008-2009 has been presented on the Faculty website together with the whole teaching materials. A course overview describes the goals of the course and suggests some prerequisites. The online tutorials offer i) a full set of course slides (in power point extension), ii) full set of teaching material offered during the course, iii) on-line exercises, iv) downloadable documents, v) links to websites of interest and downloadable software, vi) access to on-line tools for Italian text processing.
On-line availability of teaching materials provide not only the possibility of remote on-line access to actual course materials, but also appear to supply the lack of offer from traditional Italian publishing houses. The reference in the course program to only 5 chapters of *“Testo e Computer” Lenci, Montemagni, Pirrelli – Carocci Editor (2005)* is a clear evidence of the difficulty of Italian publishing houses of addressing the general public need for a more aware use of current technologies, and of their traditional and conservative attitude.

3.2 “Electronic Archive of Carlo Emilio Gadda's Works”

As a second case study we counted the website accesses to the whole literary production of Gadda (1893-1973) [http://www.ilc.cnr.it/CEG/]. See figure 1.
The creation of the "Electronic Archive of Carlo Emilio Gadda's Works" was started in 1994, when publisher Garzanti made their published text materials (Gadda, 1988-93), already prepared for photo-composition and for study purposes only, available to the ILC.
The first version of the Corpus, running in DBT 3.0, was presented on November 14, 1997 (the 104th anniversary of Gadda's birthday) at the CNR in Rome.
While the community of Gadda scholars is starting to demand something that cannot be distributed but is available for consultation at ILC - concordances -, the project started to: (i) produce lexicographic tool standards by means of an automatic processing system; (ii) apply innovative methodologies onto sample subsystems.
The web site advertised the history of the project and made paper publications available online. Linguistic resources available:
{1)} Hapax Legomena Inverse Index
{2)} L'accentazione in Gadda [Accentuation in Gadda]
{3)} Un primo censimento di termini gaddiani [A first census of Gaddian terms]
{4)} Il latino in Gadda [The use of Latin in Gadda]
{5)} Annotazioni su composti in -cola [Annotations on compounds terminating in -cola]
a) Concordances by form of La cognizione del dolore
b) Concordances by form of Pasticciaccio
c) Complete concordances of the question mark (contexts with right-hand arrangement by ?)
d) Complete concordances of the question mark (contexts with left-hand arrangement by ?)
e) Complete concordances of the exclamation mark (contexts with left-hand arrangement by !)
f) Co-occurrences of Giornale di guerra e prigionia
g) Index Locorum of Latin forms in Gadda
h) Latin forms in Horace and Gadda - Comparison table
i) Comparisons between the two versions of Pasticciaccio
j) Gaddian iterations.

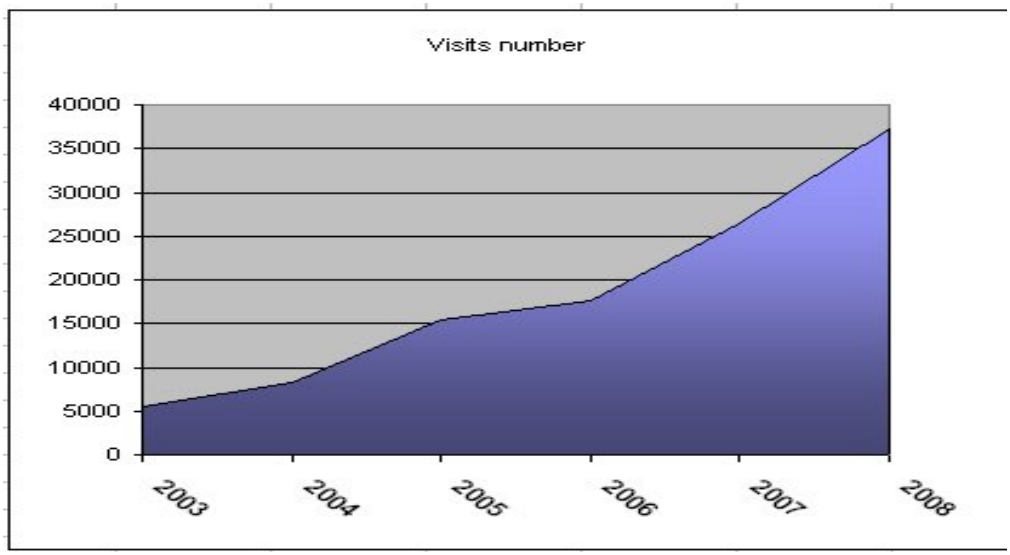


Figure 1 Visit numbers in year sequence

4. Concluding Remarks

On-line availability of tutorials and teaching materials on the web pages of Computational Linguistics courses in Italy is playing a two-fold role: a) supply a persisting lack of offer of Italian manuals from traditional publishers; b) provide a jumping-off point to general public internet users for a more aware and selective access to the web haystack, based on integration of personalized language-driven queries and precompiled information repositories and ontologies. In the near future, this is bound to change habits, demands and expectations of a new generation of self-taught web users, capable of exploiting the web in creative ways and pacing the development of novel technologies for information access.

References

Barsalou, L. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22, 577–609.

Buitelaar, P., Cimiano, P., and B. Magnini (2005). Ontology Learning from Text: an Overview. In Buitelaar et al. (eds.), Ontology *Learning from Text: Methods, Evaluation and Applications* (Volume 123 Frontiers in Artificial Intelligence and Applications): 3–12.

Bybee, Joan, Hopper, Paul (eds.) (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins.

Guarino, N. (1998). Formal Ontology in Information Systems. *Proceedings of FOIS'98*. Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

Kilgarriff A., Grefenstette G. (2003). Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, 23 (3), 333-47.

F. W. Lancaster (2003). *Indexing and abstracting in theory and practise*. Third edition. London, Facet.

Lawrence S., Lee Giles C. (1999). Accessibility of Information on the Web, *Nature*, 400, 107-9.

Manning, Chris, Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press US

Mitchell, Tom M. (1995). *Machine Learning*. McGraw Hill US.

Ramos de Carvalho E.M. (2001). Grey Literature and its contribution to knowledge Society. *67th IFLA Council and General Conference*, 1-11.

https://doi.org/10.26069/greynet-2026-000.559-tgj

Open Access to Grey Literature on e-Infrastructures: The BELIEF Project *

Stefania Biagioni, Donatella Castelli, and Franco Zoppi *(Italy)*

Abstract

The BELIEF (Bringing Europe's eLectronic Infrastructures to Expanding Frontiers) Project is a Coordination Action funded by the European Commission in the context of the FP6 and FP7 Programmes. It aims to create a platform where e-Infrastructures providers and users can collaborate and exchange knowledge. This will help ensuring that e-Infrastructures are deployed and effectively used worldwide, filling the gap separating the Research Infrastructure providers from the users, and thus contribute to the emergence of a competitive knowledge-based economy. To create this synergy among multi-disciplinary Research Infrastructure communities, BELIEF created a one-stop-shop for e-Infrastructures communities providing a Community Portal http://www.beliefproject.org/ and a Digital Library (DL) http://belief-dl.research-infrastructures.eu/ with a huge number of e-Infrastructures open access documents. Moreover, it has organised events including brainstorming, networking workshops and international conferences and publications, since BELIEF's values are firmly rooted in international cooperation with the emerging economies, particularly in Latin America, India and South Africa. The BELIEF DL – implemented on top of the OpenDLib Software System – offers uniform access to multimedia documentation and especially to grey literature (e.g. presentations, videos, technical reports, manuals, on-line tutorials, etc.), providing continuously updated information on e-Infrastructures-related projects, initiatives and events. The contents are harvested from different sources, such as projects web sites, repositories and databases. The DL provides services to support the submission, description, searching, browsing, access, preservation and visualization of these multimedia documents. Although designed to meet the needs of a specific scientific community, the technology adopted by BELIEF can be easily adapted to meet the information and collaborative needs of other user communities. The BELIEF e-Infrastructures Community ranges now on more than 80 projects, initiatives and organisations inside and outside Europe, and the DL offers nearly 15.000 documents harvested from their repositories and websites. The BELIEF Consortium is composed by Brunel University (UK), CNR-ISTI (Italy), ERNET (India), Escola Politécnica da Universidade de Sao Paolo (Brasil), Meraka Institute (South Africa), Metaware SpA (Italy), and the National Kapodistrian University of Athens (Greece).

1 Introduction

The BELIEF (Bringing Europe's eLectronic Infrastructures to Expanding Frontiers) Project aims to create an effective open workspace where e-Infrastructures providers and users can collaborate and exchange knowledge, ensuring the development and adoption of e-Infrastructures on a worldwide scale. The BELIEF DL play a key part in the project, bringing a range of benefits to e-Infrastructures stakeholders across the globe by facilitating the exchange of knowledge and experiences through a single and easily accessible tool. The BELIEF Project arose from the awareness that a gap existed between Research Infrastructure providers and users. In order to bridge this gap, a complete and common source of information on e-Infrastructures was needed, both for users demanding provision and resources and for providers intending to extend their user base and develop their systems. The BELIEF DL responds to this demand by providing users with documentation matching their search criteria accurately and according to their interests and professional profile. This paper focuses on the implementation of this key component.

General outcomes of the design phase are reported in [7]. In the following (Section 2) the main characteristics of the DL are described. General concepts regarding the implementation of the DL are outlined (Paragraph 2.1). These general concepts represent the logical link between basic requirements and the components of the implemented solution (Paragraph 2.2). Then (Paragraph 2.3) the organization of the User Interface is introduced.

* First published in the GL11 Conference Proceedings, March 2010

Section 3 summarizes the most relevant usage data as per the statistics gathered during the whole projects lifetime, from 2006 onward.

This paper won't cover all of the relevant topics of the implementation work, for those please refer to [4]. Hereafter are summarized the ones covered in this paper (Section 4):

- Definition of the metadata structure implemented by the DL.
- Definition and implementation of protocols and tools for the submission of metadata and documents to the DL.
- OAI-PMH Compliance.
- Integration with external portals.

Conclusions and results achieved are finally outlined in Section 5.

2 The Digital Library

The DL was designed to serve the needs of e-Infrastructures research and industrial users that want to keep up to date with existing projects and the latest developments in e-Infrastructures. The DL offers its user community advanced services to uniformly access multimedia documents such as technical reports, presentations, videos, manuals, on-line tutorials, etc. These documents contain the very latest details on e-Infrastructures related projects, initiatives and events.

The material maintained in the DL is harvested from different sources, such as web sites, repositories and databases of e-Infrastructures Projects, Initiatives and Organisations. The DL organises the harvested information according to the information needs of the user communities rather than according to its physical format, structuring and distribution on diverse sources. This means that it is capable of providing users with multiple virtual views of the existing documentation. To this end, an extremely accurate collection and analysis of the requirements of potential users was made before the DL was created.

The DL provides services to support the submission, description, searching, browsing, retrieval, access, preservation and visualization of multimedia documents. Users can define the information space which they want to search/browse upon in terms of collections (i.e. sets of documents) selected from those managed by the DL. Collections can be created interactively, based on the archives the documents are to be selected from. Different search/browse options are offered: Google-like or fielded (with fields selected from a variety of known metadata formats), with or without relevance feedback. Users can search/browse any information associated with digital documents and their parts.

2.1 Preliminary Design - Methodology

A design phase based on a deep analysis of the Project's target audience needs was carried out with a selected set of Research Infrastructure Entities (Projects, Organisations and Initiatives). The methodology adopted for requirements collection was basically structured in the following activities:

- Drawing up a Questionnaire to collect requirements on functional requirements, documents to be managed and their related metadata (descriptive data).
- Identifying proper interfaces within Entities, contacting them, obtaining and discussing requirements.
- Matching similar requirements from diverse Entities.
- Verifying the quality of the metadata and making the semantics and use of metadata conform.
- Verifying the quality of documents, analysing the types of documents used by diverse Entities with different semantics and making them conform.
- Drawing up a Memorandum of Understanding to be signed with each of the Entities collaborating in subsequent releases to ensure clear, comprehensive and effective interaction and collaboration.

2.2 DL's General Concepts

This section introduces the main concepts characterising the DL from the perspective of its end-users. These concepts represent the logical link between the basic users' requirements and the components of the solution implemented, as follows:

Users of the BELIEF **DL** access the DL's **Resources** according to their **Role** and to the **Access Rights** they hold on **Collections**. These are grouped in user-specific **Information**

Spaces. A **User Profile** is then associated with each User, recording his/her characteristics in terms of Role and Access Rights.

Where:

Users Are the actors entitled to interact with the DL. Different types of Users can be provided.

DL From a high-level perspective, a DL is an organisation capable of collecting, creating, accessing and managing Resources.

Resources The resources that can be accessed by Users, depending on their Role and Rights. Diverse types of Resources can be provided (**Content** and **Services**)
Resources of type Content are gathered from **Information Sources**.

Information Source This term is used to indicate a repository belonging to an Entity (Project, Organisation or Initiative) from which metadata and documents are collected to form the DL. Information Sources are characterised by the type, structure and semantics of the metadata and documents they contain.

Role A function within the context of an organisation (in our case the BELIEF DL) with some associated semantics regarding the authority and responsibility conferred to the user. Different types of Roles can be provided (**Consumers**, **Providers** and **Librarians**).

Access Rights The actions allowed on Resources (Create, Modify, Delete, Search/Browse, Use/Perform, etc.).

Collection A set of Resources of type Content defined according to characterisation criteria. Collections represent the mechanism to organise Content in order to provide focused views of the Information Space, and allow DL end-users to access to thematic parts of the whole Information Space.

Information Space Corresponds to the **set of Collections of the DL** that each user can access. It defines the scope of Content against which the Browse/Search operation are executed. The Information Space can be dynamically modified by the user to restrict/enlarge such scope.

User Profile Contains information about the user. It may be viewed as a set of characteristics with associated values. Each piece of user information alone and in conjunction with others determines system behaviour towards the user.

2.3 General Characteristics of the Implemented Solution

The BELIEF DL is created as an instance of the OpenDLib [4] Software System, thus being capable of offering the following services to its users:

- Creation, submission, search, browse, access, and preservation of multimedia documents.
- Definition of the their personal Information Space which they want to search/browse upon in terms of collections selected from those managed by the DL. Collections can be interactively created defining:
  - the condition that is satisfied by the members of the collection;
  - which archives the documents are to be selected from.
- Different search/browse options: Google-like or fielded (with metadata elements selected from a variety of known metadata formats). Users can search/browse any information associated with digital objects and their parts. As a result of their search/browse operations, users obtain a set of result pages with the list of digital objects that satisfy their request. By clicking on an object, users can access any of its multiple manifestations. In particular, they can select the one that is compatible with the software installed on the computer they are working with.
- Full compliant Open Access via the OAI-PMH protocol.

In the implementation of the DL particular effort has been devoted to the following issues:

- Implementation of harvesters supporting standard formats and specific conventions on Information Sources side. Presently, the following standard protocols and coding formats are supported by the harvesting tools of the DL:
  - Metadata encoding formats: DC, DCQ (recommended), MARC, UNIMARC, MARC21, MARCXML.

- Metadata harvesting protocols: OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting (recommended) or any API call returning an XML file containing metadata encoded in one of the above mentioned formats.
- File formats: XML (recommended), RSS.

- Implementation of "document models" and of a web-based interface allowing users to easily submit metadata and documents to the DL using those models.
- Implementation of a set of APIs to be easily used by programmers to interface basic functions of the DL.
- Harmonization of concepts and practices (e.g. use of metadata and terms and of different types of document) to be benefited by the whole Community in facilitating knowledge communication and document exchange.
- Definition and/or homogenisation of vocabularies used both for metadata and document content description, categorization and search.

The benefits coming from the use of the DL are twofold:

- On the Information Provider's side:
  - e-Infrastructures projects' documents can be accessed by a wider users' community.
  - This wider audience is also a more targeted audience. The DL provides users with the most accurate selection for their information needs, in a way that bridges them directly with the projects' material.
  - It gives projects' audience a one-stop-shop to find the precise documentation that they require about the projects, thus it is a valuable target to send inquirers to, should they be searching quickly for projects' documentation.
  - Documents are accessible via different formats.
  - The DL supports document preservation.
  - Projects can submit their documents directly to the DL. BELIEF offers a submission interface and a document hosting service. To ensure that any new documents can be easily collected by the DL a standard method is provided to describe them.
- On the Information Consumer's side:
  - While accessing a website and browsing a project's material, users get material from that project/website only, using the DL they get the relevant information from many projects' sites in a single shot. This give them a wider range of information on e-Infrastructures, how they work, how to use them, etc..
  - User can search documents in a uniform way, accessing via a unique interface documents produced by different providers.
  - The material accessible is extremely well focused on the users' needs. This is not like Google which may often come up with nothing or thousands of results or it yields material that is not relevant or up to date.

2.4 The User Interface and its usage

The DL's User Interface (UI) has been designed to reflect the most recent advances in UI usability:

- The look and feel let users have an extremely comfortable access to functionality and content.
- The overall navigation structure has been designed to minimise the number of clicks the users need to access any content.
- All most common functions can be easily accessed via one-click commands.
- Most of the relevant information related to a document are shown in the same window.
- A complete On-line User Guide has been implemented by means of a friendly Wiki capability.

The overall work area of the DL (the *DL Desktop*) is clearly organised into three sections [2]:

- On the left side is the "Community" section. According to each specific account permissions – after having executed a Login – a number of functions are accessible here: "News", "Personal Profile", "Users & Groups Management", "Documents Management", and "Information Space Management".
- On the right side is the "Content Access" section, where the content access functions (Browse, Search) are accessible. A "News" section is also shown, presenting relevant information about the DL from different points of view (events, user services, technical, etc.).

- In the middle is the "Information Space" section of the Desktop where available Collections are listed and the content resulting from Browse and Search operations is shown.

The following figures show respectively the overall look of the UI, the definition of a search criteria and the way for accessing the actual content of a document – once it has been retrieved via a Browse/Search operation. Please refer to [3] and [6] for details, or access directly [2].

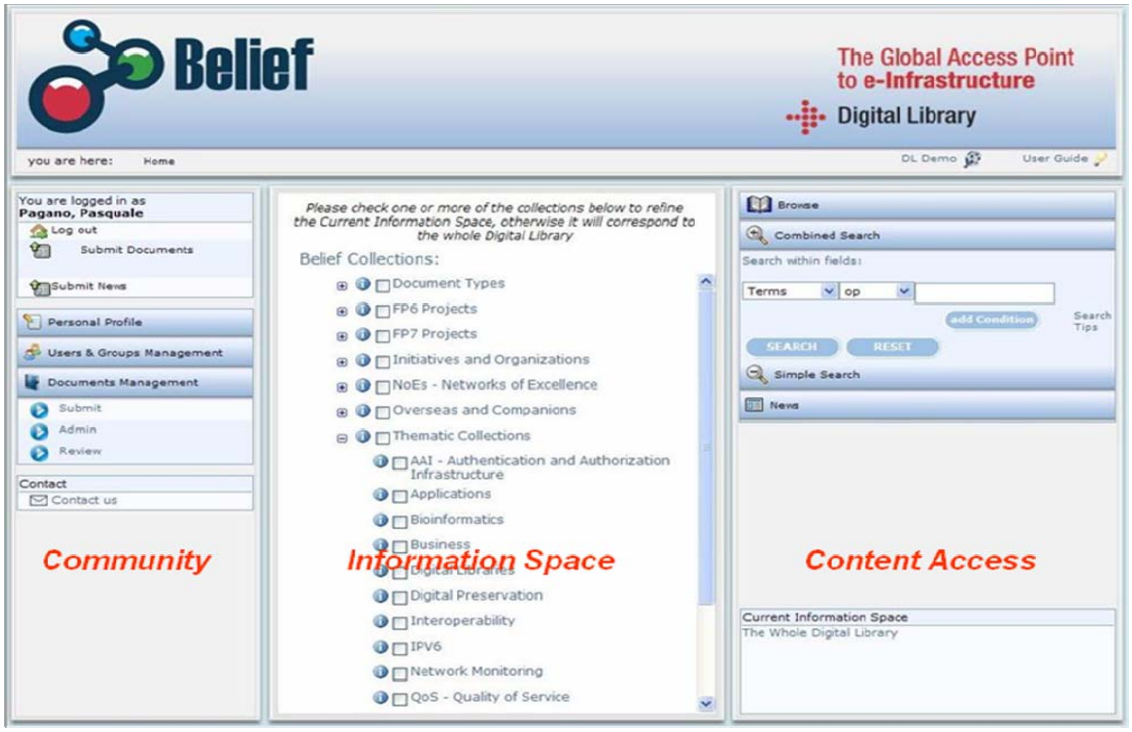


Figure 1 – The DL Desktop organization

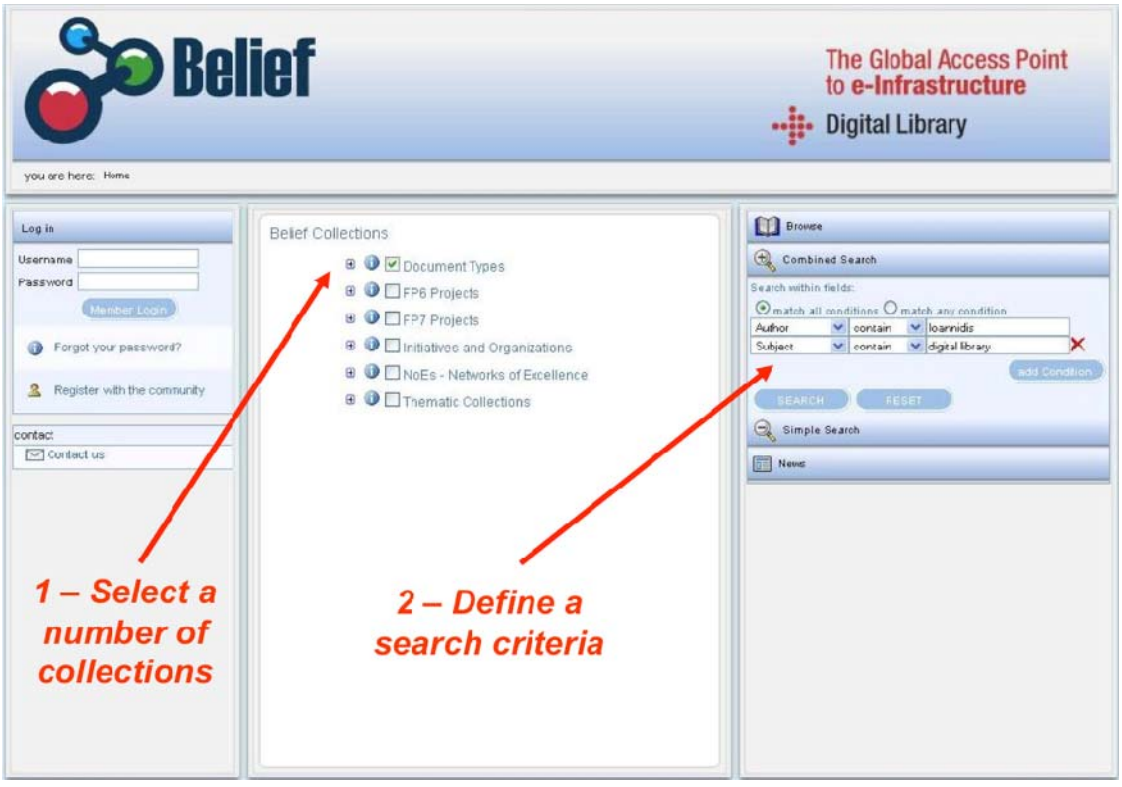


Figure 2 – Searching content

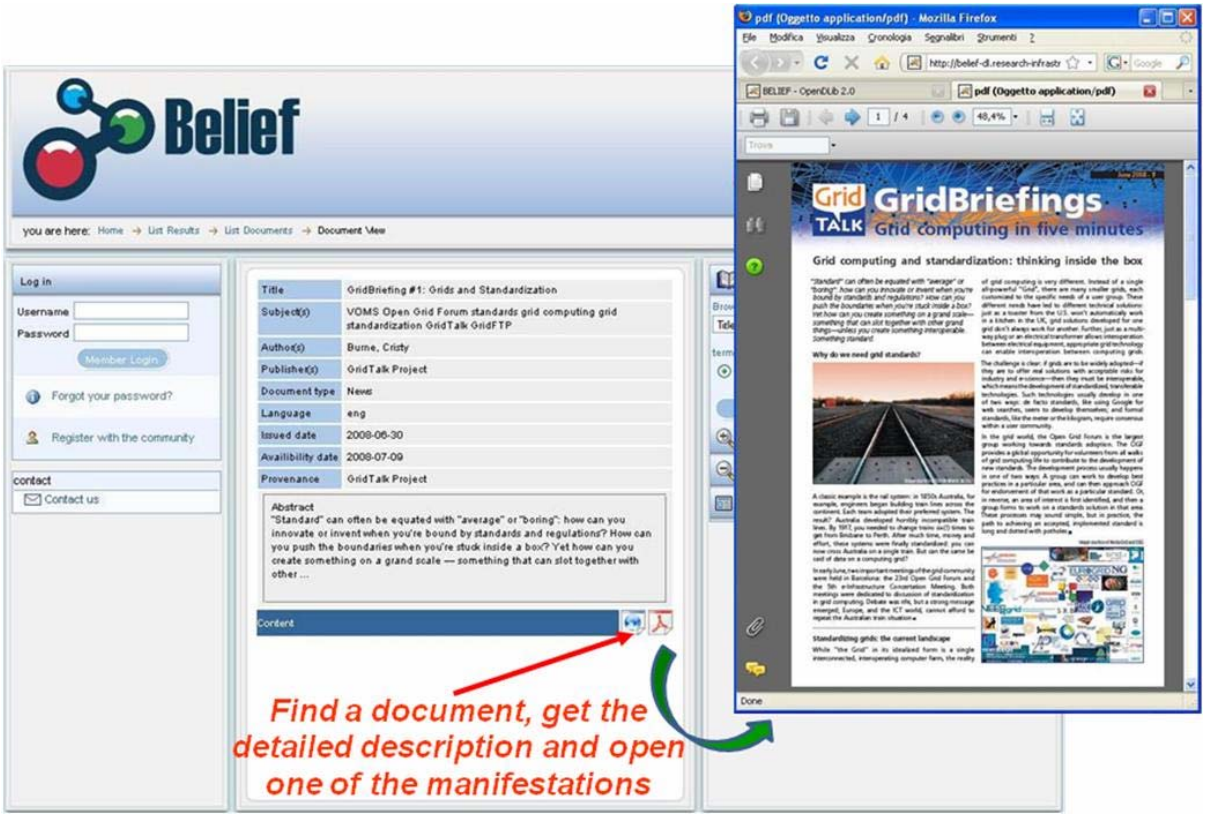


Figure 3 – Accessing content

3 BELIEF Events and Community's DL growth and usage

BELIEF successfully organised a series of events which broke new ground in the arena of Grid-empowered infrastructures and e-Infrastructures. These events brought together Grid & e-Infrastructures experts and technology developers, IT innovators from both enterprise and research, decision-makers, and scientific policy-makers. Participants came together to share ideas and knowledge, discuss how technological challenges could be tackled, strengthen alliances between business and research, and help unleash the potential of e-Infrastructures.

All of these international events have helped to setup a synergetic network of relationships and to valuably increase the content of the Digital Library with material submitted by all Community's Entities. Thanks to the outcomes of these events and to the established synergies and relationships, a significant growth of the Community has been achieved, largely exceeding the original planning of the project.

The rich and differentiated content offered by the DL is reported by the following chart, with a clear prevalence of material produced by Conference and Technical Meetings, then Presentations, Articles, Training Material, Deliverables, Technical Documents, etc..

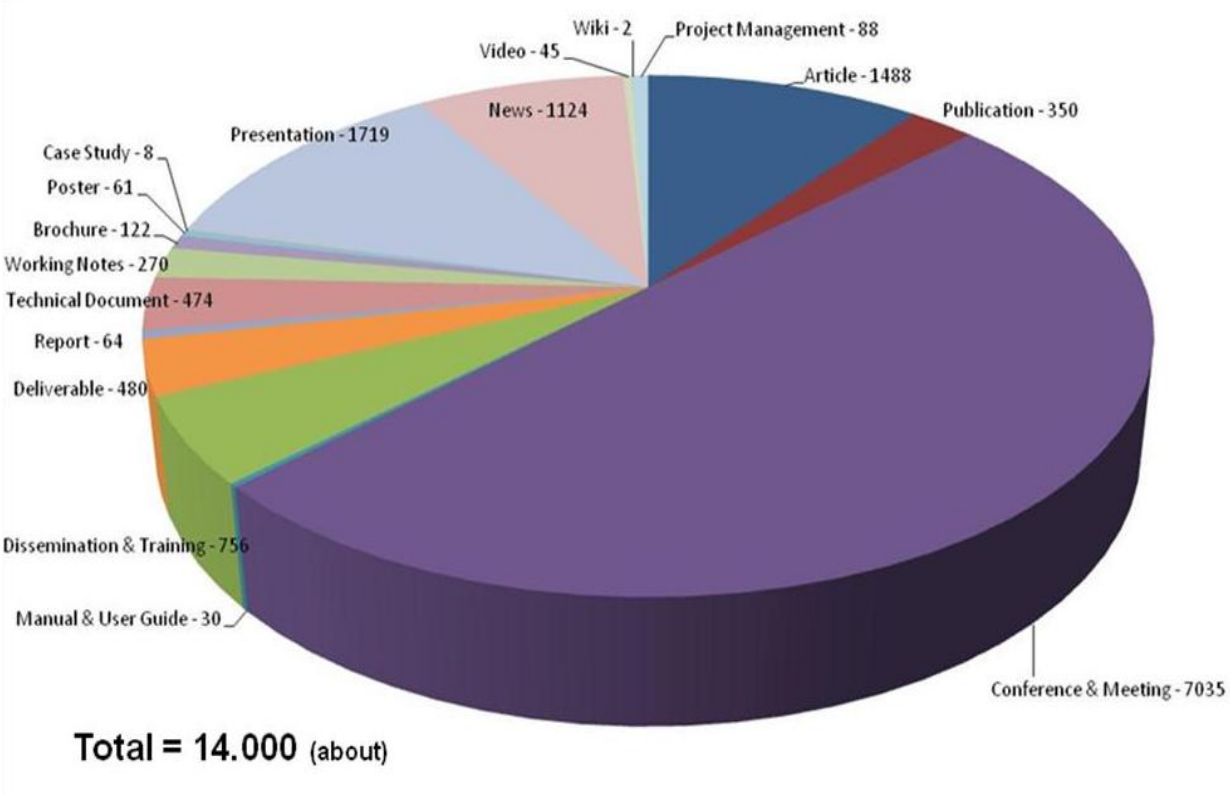


Figure 4 – DL's content type

It's meaningful to point out the trend of accesses as logged by the DL's statistical tools (see following figure). The chart clearly shows a number of peaks related – on one side – to the aforementioned events and – on the other side – to key dates corresponding to milestones in projects' lifetime (reviews, call for proposals, etc.). The decrease in the last two months is related to the particular conjunction of the year's period and of most of projects' lifecycle period.

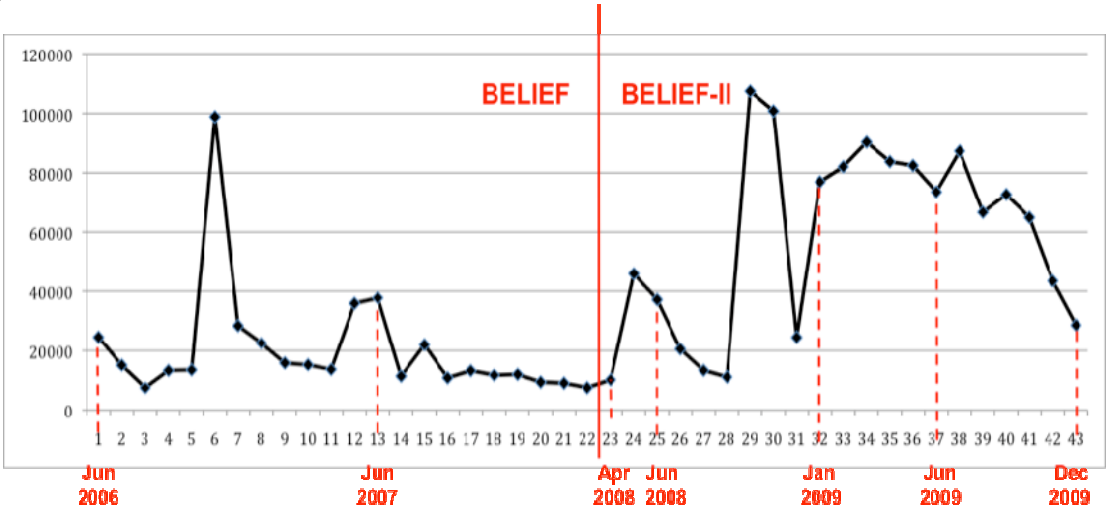


Figure 5 – DL hits per month from June 2006 to December 2009

It's worthwhile also mentioning that both the average and the maximum number of hits registered a growth of one order of magnitude from the first phase of the project (marked as "BELIEF" in the previous figure) to the second one (marked as BELIEF-II).

The detailed analysis of the users' provenance evidenced that among the **Top Sites** (excluding generic name-servers, and accesses coming from known EU projects and projects' partner sites) from which the DL is accessed the following should be mentioned:

- Google, with the GoogleBot web spider
- Yahoo, with its Web Crawler

- A number of telecoms' portals via their search services, e.g. Fastweb, Telecom Italia, Vodafone, On Telecoms, etc.
- A number of search engines, e.g. MSN Search, Cuil, Dotbot, etc.

It's worth notice that these data, joined to the info related to the provenance of accesses, highlight the wide spreading and variety of consumers achieved by the DL.

Finally, the **Top Operations** as registered by the statistical tools are the following:

- Query (simple)
- Browse
- OAI access
- Query (complex)
- Submission of new content

These data manifestly show that the most frequent operations on the DL are queries performed on both all Metadata Elements and Full Text, followed by generic browse of the content and by accesses from sites/repositories making use of the OAI/PMH protocol to query the DL. Query focused on specific Metadata Elements and the insertion of new content in the DL follow, as expected to be.

4 Technical insights

Albeit referring to [4] for a complete documentation of the technical characteristics of the OpenDLib Software System, this section briefly summarises the most relevant aspects of the BELIEF implementation, namely:

- Definition of the metadata structure implemented by the DL.
- Definition and implementation of protocols and tools for the submission of metadata and documents to the DL.
- Integration with external portals.

4.1 DL Metadata

With OpenDLib, resources can be catalogued with multiple metadata formats. The BELIEF DL uses Dublin Core Qualified (DCQ) encoding for the purpose of interoperability since DCQ enables the enhanced sharing of information between Information Sources adopting different coding with no loss of semantics [8].

A total of 17 DCQ metadata are currently supported by the BELIEF DL (The metadata in bold have to be considered mandatory for an effective classification in the DL): **Title**, **Creator**, **Subject**, **Description**, **Publisher**, Contributor, **Date**, **Type**, **Status**, **Format**, **Identifier**, **Source**, **Language**, Relation, Coverage, Rights, and **Provenance**.

The detailed description of the semantics and usage of such metadata elements is given in [5]. Implemented or suggested controlled vocabularies are introduced for Type, Status, Format, and Language, whilst the definition of controlled vocabularies has been undertaken for Creator and Subject and will be carried out along with the implementation of an "Authority File Control Service" to support librarians and administrators in such arduous task.

4.2 Submission of Metadata and Documents

Different methods for submitting documents and metadata to the DL are supported:

- On-line submission via Document Models (i.e. Web forms).
  This can be done either using the native DL facility (accessible for authorized users through the "Community" section of the DL Desktop) or via integration with external portals (e.g. BELIEF itself and D4Science, as shown in the following section).
  In the former, the document models and the underlying metadata structure are designed and implemented by the DL's Librarian. Different pre-defined models are supplied to cope with the "standard user" needs. In addition, a "Free Model" is supplied, which can be modified by skilled users according to their specific needs.
  In the latter, the models and the metadata are designed in cooperation with the administrators of the external portals, to exactly fit with their requirements.
- Harvesting from existing Information Sources, implemented by specific modules and interfaces.

To this end, a number of different harvester has been implemented for repositories supporting any kind of programmatic access. As to the formats, both DC/DCQ and different "flavors" of MARC have been used, whilst both OAI-PMH and a number of proprietary interfaces are supported to implement the communication protocol. In some relevant case, even HTML parsers have been implemented to access repositories not supporting any of the above.

- Batch submission via an XML schema based file.
  Finally, bulk load from repositories not supporting any on-line access facility has been implemented via batch processing of XML files. These are based on an XML schema supplied by the DL's Librarian and fine-tuned with the administrators of those repositories.

4.3 OAI-PMH Compliance

The DL features an OAI-PMH [9] fully compliant implementation which let the DL provide effective tools to support the Principles of the European Commission's Communication of Scientific Information [10] and the subsequent Competitiveness Council Conclusions on Scientific Information in the Digital Age [11].

These tools will also promote widespread adoption of an Open Access Policy that will lead towards global and seamless dissemination of publicly-funded research results (publications and data), as set out by the ERC Scientific Council Guidelines for Open Access [12] and the Open Access Pilot [13] launched by the European Commission.

To apply all of these, the integration of the DL within the DRIVER Infrastructure [14] was undertaken, implementing the full compliance with the DRIVER Guidelines 2.0 released by the DRIVER Consortium. DRIVER is the well known European data infrastructure connecting hundreds of digital repositories of institutions and research organisations. Such integration produced a first implementation which is now going to be refined with a close interaction with the DRIVER Consortium.

4.4 Integration with External Portals

An advanced and powerful characteristic of the BELIEF DL is its capability of being integrated in an external portal, that is the portal implemented by another project, willing to strictly interact with the DL to let this behave like a component/service of the portal itself.

Such integration can be achieved via either the use of the APIs provided by the underlying OpenDLib System, or by incapsulating specific code modules – especially developed for the BELIEF DL – within the external portal.

The former approach was implemented in the BELIEF Portal, which interact with the DL by means of APIs like "List Authorities" (i.e. the Entities authorized to submit content), "Submit Document", etc.. The latter was used in the D4Science project's portal, since this required the DL to run in a sort of "sand box" within the Portal itself.

The following figures show the implementation of the D4Science project, where the authentication of the user and the submission of a document to the DL are embedded in the portal. In a similar way a "plug-in-search" capability was integrated in the portal to implement a direct access to the BELIEF DL content.

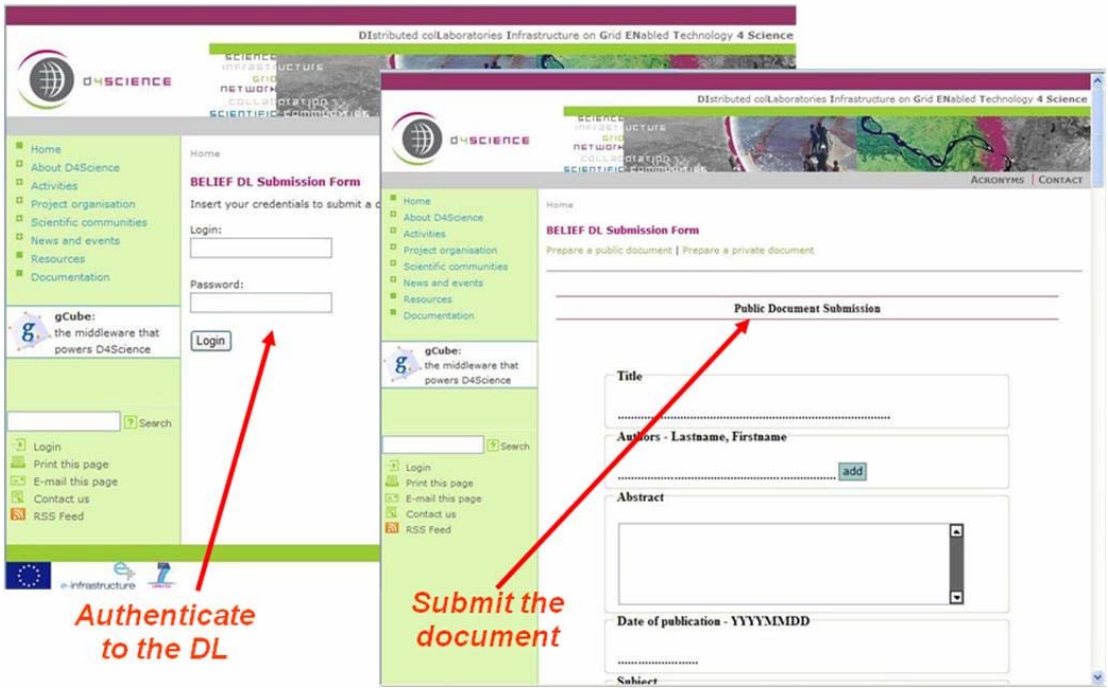


Figure 6 – DL integration with D4Science (Authentication and Submission)

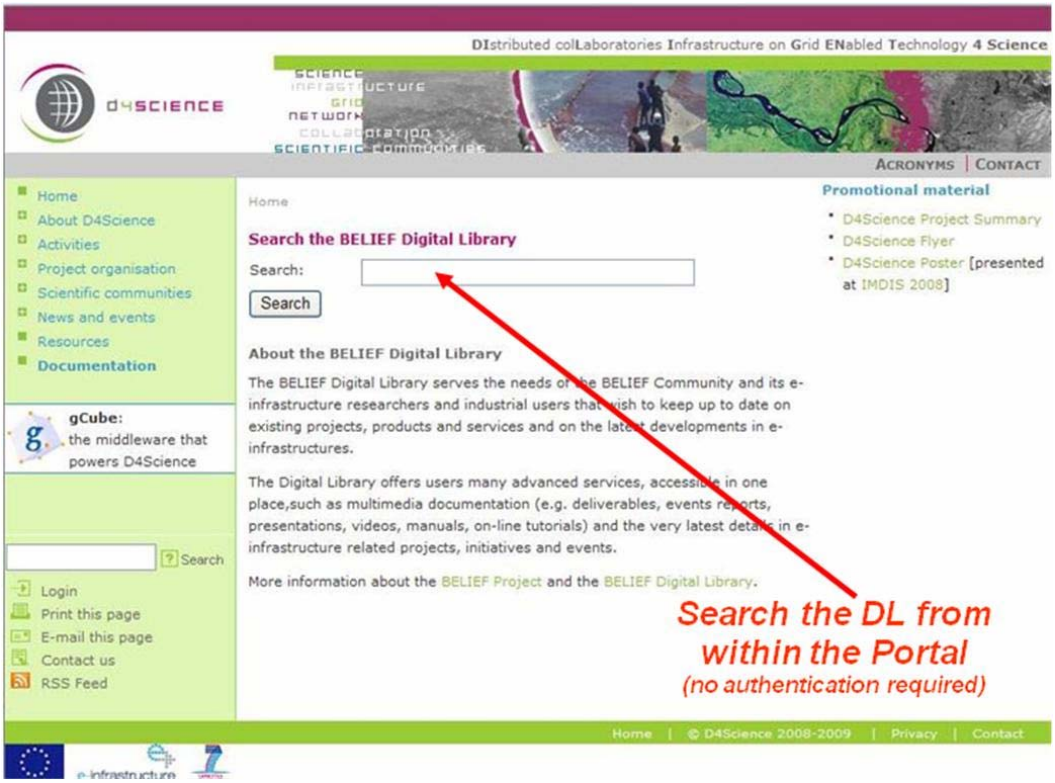


Figure 7 – DL integration with D4Science (Search)

5 Conclusions and Future Work

The effective use of a Digital Library within the scope of the BELIEF Coordination Action was shown. Specific implementation issues were highlighted, as well the conceptual aspects characterising this application context.

Focus was put on harmonisation of metadata coming from different Information Sources and on harvesting rules, protocols and formats as well as on specific harvesting tools implemented to cope with the diverse characteristics of those Sources. Compliance to open

standards as well openness of the DL's architecture was introduced, with special emphasis on the DRIVER and D4Science examples respectively.

Finally, importance of a proper and controlled usage of metadata elements and values was discussed. Implemented or suggested controlled vocabularies were introduced, whilst the implementation of an "Authority File Control Service" was envisaged. Its detailed design and implementation will be one of the main goal of the project.

References

[1] BELIEF Project http://www.beliefproject.org

[2] BELIEF Digital Library http://belief-dl.research-infrastructures.eu/

[3] BELIEF Digital Library Wiki User Guide https://userguide.wiki.belief.research-infrastructures.eu/

[4] OpenDLib website http://opendlib.research-infrastructures.eu/

[5] Zoppi F., Biagini F. *BELIEF-II D3.4.1 DL - Release Notes*. Bringing Europe's eLectronic Infrastructures to Expanding Frontiers - Phase II (BELIEF-II) - FP7-Infrastructures-2007-2 - 223759. Deliverable BELIEF-II D3.4.1, 2008

[6] Zoppi F., Baldacci M. B., Biagini F. *BELIEF-II D3.3.1 DL - User Guide*. Bringing Europe's eLectronic Infrastructures to Expanding Frontiers - Phase II (BELIEF-II) - FP7-Infrastructures-2007-2 - 223759. Deliverable BELIEF-II D3.3.1, 2008.

[7] Zoppi F., Artini M., Baldacci M. B., Biagini F. - *BELIEF - WP1D5.4 - Bridging Information Consumers & Providers*. BELIEF (Bringing Europe's eLectronic Infrastructures to Expanding Frontiers) - FP6-2004-Infastructures-6 - 026500. Deliverable BELIEF WP1D5.4, 2008.

[8] The Dublin Core Metadata Initiative Open Forum http://www.dublincore.org/

[9] The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html

[10] Principles of the European Commission's Communication of Scientific Information http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf

[11] Competitiveness Council Conclusions on Scientific Information in the Digital Age http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf

[12] ERC Scientific Council Guidelines for Open Access http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf

[13] Open Access Pilot ftp://ftp.cordis.europa.eu/pub/fp7/docs/open-access-pilot_en.pdf

[14] DRIVER - Digital Repository Infrastructure Vision for European Research http://www.driver-repository.eu/ http://www.driver-community.eu/

https://doi.org/10.26069/greynet-2026-000.559-tgi

Twelfth International Conference on Grey Literature

National Technical Library, Prague, Czech Republic, 6-7 December 2010

Transparency in Grey Literature
Grey Tech Approaches to High Tech Issues

Conference Program

DAY ONE

9:00-10:30

Opening Session

Chair, Welcome Address	***Martin Svoboda, Director National Technical Library, Czech Republic***
Opening Address	***Dr. Radim Polčák, Head Institute of Law and Technology, Masaryk University, CZ***
Keynote Address	***Blane K. Dessy, Executive Director FLICC-FEDLINK, Library of Congress, USA***

11:00-12:30

Session One – Redefining Grey Literature

Chair, *Christiane Stock, Institute for Scientific and Technical Information, INIST/CNRS, France*

Towards a Prague Definition of Grey Literature *Joachim Schöpfel, University of Lille, France*

A terminology-based Re-definition of Grey Literature *Gabriella Pardelli, Manuela Sassi, and Claudia Marzi, Institute of Computational Linguistics, National Research Council, Italy*

Peering through the Review Process: Towards Transparency in Grey Literature *Dominic J. Farace, Grey Literature Network Service, Netherlands*

An Analysis of Current Grey Literature Document Typology *Petra Pejšová, National Technical Library, Czech Republic and Marcus Vaska, Health Information Network Calgary, Canada*

13:30-15:00

Session Two – New Stakeholders in Grey Literature

Chair, *Elizabeth Newbold, The British Library, United Kingdom*

Global Grey Literature in Health: Identification, Preservation, and Sustainability *Janie Kaplan, Lea Myohanen, and Elizabeth Taylor, New York Academy of Medicine Library, United States*

The Comparative Study of Grey Literature Organisation and Approach: Two Countries, Similar and Different *Primož Južnič, University of Ljubljana, Slovenia; Petra Myšková and Richard Papik, Charles University, Czech Republic*

Using Environmental Grey Literature to Engage Public Participation in Decision Making: Government Transparency in Coastal Policy Development *Bertrum MacDonald, Suzuette Soomai, and Peter Wells, Dalhousie University, Canada*

Grey Literature Past and Present – the Evolution of Library Annual Reports *Susan E. Searing, University of Illinois, United States*

15:30-16:30

Introductions to Poster Presentations

Chair, *Petra Pejšová, National Technical Library, Czech Republic* **(See further Listing of Posters, below)**

17:00-19:00

Inaugural Reception

Special Conference Welcome

GL12 Program and Conference Bureau
TextRelease
Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com ▪ **conference@textrelease.com**
Tel/Fax +31 (0) 20-331.2420

https://doi.org/10.26069/greynet-2026-000.559-tgi

Twelfth International Conference on Grey Literature

National Technical Library, Prague, Czech Republic, 6-7 December 2010

Transparency in Grey Literature

Grey Tech Approaches to High Tech Issues

Conference Program

DAY TWO

09:00-10:30

PARALLEL SESSIONS:

Sponsor Showcase and Poster Presentations **(See further Listing of Poster Titles and Authors, below)**

10:30-12:00

SESSION THREE – STANDARDIZATION IN GREY LITERATURE

Chair, *Janie Kaplan, New York Academy of Medicine, United States*

From OpenSIGLE to OpenGrey : Changes and Continuity *Christiane Stock and Nathalie Henrot, Institute for Scientific and Technical Information, INIST-CNRS, France*

Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository *Jindřich Mynarz and Ctibor Škuta, National Technical Library, Czech Republic*

Current Status and Issues in Collection and Dissemination of National Technical Reports in Korea - Improvement through the use of the National Science and Technology Information Service (NTIS) *Jae-Soo Kim, Hee-Yoon Choi, Inseok Song, Kiseok Choi and Sung-Jin Jhun, KISTI, Korea*

A Profile of Italian Working Papers in RePEc *Rosa Di Cesare, Daniela Luzi, Marta Ricci and Roberta Ruggieri Institute of Research on Population and Social Policies, IRPPS/CNR, Italy*

13:00-14:30

SESSION FOUR – NEW FRONTIERS IN GREY LITERATURE

Chair, *Joachim Schöpfel, University of Lille, France*

Scientific Data: Increasing Transparency and Reducing the Grey *Bonnie C. Carroll and June Crowe Information International Associates Inc., United States*

Developing small worlds of e-science: using quantum mechanics, biological science, and oceanography for education and outreach strategies for engaging research communities within a university *Plato L. Smith II, Florida State University, United States*

Invenio: A Modern Digital Library System for Grey Literature *Jérôme Caffaro and Samuele Kaplun, European Organization for Nuclear Research, CERN, Switzerland*

GL Transparency: Through a Glass, Clearly *Keith G. Jeffery, Science and Technology Facilities Council, United Kingdom and Anne Asserson, University of Bergen, Norway*

14:45-15:30

CLOSING SESSION - REPORTS FROM CHAIRPERSONS, CONFERENCE HANDOFF, AND FAREWELL

Chair, *Martin Svoboda and Petra Pejsova, National Technical Library, Czech Republic*
Dominic Farace, Grey Literature Network Service, Netherlands

15:30-16:30

POST-CONFERENCE TOUR OF THE NATIONAL TECHNICAL LIBRARY

GL12 Program and Conference Bureau

TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com ▪ **conference@textrelease.com**
Tel/Fax +31 (0) 20-331.2420

https://doi.org/10.26069/greynet-2026-000.559-tgi

Twelfth International Conference on Grey Literature

National Technical Library, Prague, Czech Republic, 6-7 December 2010

Transparency in Grey Literature
Grey Tech Approaches to High Tech Issues

Conference Program

Poster Session

Listing appears in Alphabetical Order by 1st Author

Grey Literature in Engineering Sciences and Technology and its use pattern in the research institutions in India: The case study of the Karnataka State N. Chowdappa, BMS College of Engineering and C.P. Ramasesh, University of Mysore, India

Awareness Not Genre Matters: Findings from a Case Study of the Scientific Grey Literature of an Inter-governmental Organization Danielle Cossarini, Bertrum MacDonald, and Peter Wells, Dalhousie University, Canada

Delivering Grey Literature with a "Books by Mail" Lending Library Deg Farrelly, Arizona State University, United States

Circulation Improvement of Articles in Journals written by Non-English Language - Development of a Special Journal Titles Translation List of Journals written in Japanese for the International Bibliographical Database Mayuki Gonda, Katsuhiko Kunii, Hidemitsu Nakajima, Kiyoshi Ikeda, and Keizo Itabashi, Japan Atomic Energy Agency (JAEA), Japan

Scanning to PDFA: A Demonstration Project for Access and Preservation of Grey Literature Thomas Derek Halling, Esther Carrigan, Nancy Burford, Gail Clement, Texas A&M University; and Heather Moberly, Oklahoma State University, United States

What the Future holds in store for GreyNet International: Findings from a Business Report 2010-2011 Dominic J. Farace and Jerry Frantzen, TextRelease, Netherlands

Managing the Access to the Scientific Grey Literature through Internet and Free Software Clara Lanza, ICCOM-CNR, Italy

Publishing the vocabulary of the types of grey literature as linked data Jindřich Mynarz, Petra Pejšová, and Tereza Simandlová, National Technical Library, Czech Republic

National Repository of Grey Literature in the Czech Republic Petra Pejšová and Iveta Fürstová, National Technical Library (NTK), Czech Republic

Extending the "Facets" concept by applying NLP tools to catalog records of scientific literature Eugenio Picchi and Manuela Sassi, Institute of Computational Linguistics; Stefania Biagioni and Silvia Giannini, Institute of Information Science and Technologies, CNR, Italy

Usage assessment of an institutional repository : A case study Hélène Prost, INIST-CNRS; Isabelle Le Bescond, University of Lille 1 and Joachim Schöpfel, University of Lille 3, France

ISO Standards for Quality Education and Training for Facilitating Access to Grey Literature C.P. Ramasesh, University of Mysore and N. Chowdappa, BMS College of Engineering, India

https://doi.org/10.26069/greynet-2026-000.559-tgj

Twelfth International Conference on Grey Literature

National Technical Library, Prague, Czech Republic, 6-7 December 2010

Transparency in Grey Literature

Grey Tech Approaches to High Tech Issues

Registration Form

The Conference Fee includes attendance during the Sessions, a copy of the GL12 Program and Abstract Book, the full-text of the Conference Papers and PowerPoint Presentations, as well as the conference pouch and participant badge. Also included are lunches, refreshments during the morning and afternoon breaks, and the Conference Reception.

Payment after October 1, 2010 add a € 75 surcharge

☐ **€ 495** Participant Fee ☐ **€ 395** Author Fee ☐ **€ 275** Day Pass ☐ **€ 150** Student Fee

☐ I am a *GreyNet* Member entitled to 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 No. 3135.85.342 - Rabobank, Amsterdam
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to GL12 and Registrant's Name

❑ MasterCard/Euro ❑ 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, here _____

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.**

GL12 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

Author and Title Index 2010

Biagioni, Stefania **149**
Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in data processing. She is currently a member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione (ISTI), an institute of CNR. She is head librarian and member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been the responsible of ERCIM Technical Reference Digital Library (ETRDL) and currently of the PUMA (PUblication MAnagement) & MetaPub, http://puma.isti.cnr.it. She has co-authored a number of publications dealing with digital libraries. Her activities include integration of grey literature into library collections and web access to the library's digital resources, including electronic journals and databases. She is a member of GreyNet since 2005. stefania.biagioni@isti.cnr.it

Castelli, Donatella **149**
Donatella Castelli joined the CNR-ISTI in 1987. Since then she has participated actively in several EU and Nationally funded projects on Digital Libraries and Research Infrastructures. Among them: "ERCIM Digital Library-DELOS", DELOS ECHO, SCHOLNET (as scientific co-ordinator), CYCLADES, Open Archives Forum, DLib Competence Center, DILIGENT (as scientific co-ordinator) and DRIVER. Currently she is the Scientific Co-ordinator of the D4Science and D4Science-II projects. She is also involved in the BELIEF-II, DRIVER-II, OpenAIRE, DL.org and GRDI2020 FP7 projects. Her research interests include digital libraries content modelling, interoperability and architectures.
donatella.castelli@isti.cnr.it

Luzi, Daniela **137**
Daniela Luzi is researcher of the National Research Council at the Institute of research on populations and social politics. Her interest in Grey Literature started at the Italian national reference centre for SIGLE at the beginning of her career and continued carrying out research on GL databases, electronic information and open archives. She has always attended the International GL conferences and in 2000 she obtained an award for outstanding achievement in the field of grey literature by the Literati Club. d.luzi@irpps.cnr.it

Marzi, Claudia **145**
Claudia Marzi graduated in Foreign Languages and Literature at Pisa University in 1998, with the dissertation "The power of words: language creativity in Edgar Allan Poe's narrative". Since 2001 she has been working at the CNR Institute for Computational Linguistics in Pisa. She contributed to many national (funded by Italian Ministry of Research) and international (funded by European Commission) projects. Principal organizer of the European Science Foundation Exploratory Workshop "Words in action: Interdisciplinary approaches to understanding word processing and storage" (Pisa, Oct. 2009); coordinator of a ESF Research Networking Programme proposal (call 2009) on "The European Network on Word Structure. Cross-disciplinary approaches to understanding word structure in the languages of Europe". Member of ILC scientific board at Institute for Computational Linguistics. claudia.marzi@ilc.cnr.it

Mynarz, Jindřich **129**
Jindřich Mynarz has got a bachelor's degree in Library and information science at the Institute of Information Studies and Librarianship, Charles University in Prague, and he continues with New media studies accredited at the same university. He works at the Development of electronic services department at the National Technical Library in Prague, Czech Republic. The main focus of his work is on library data and their transformation to more web-compatible data models and their exposing in conformance with linked data principles. Other areas of interest include knowledge organisation systems, particularly in the ways in which they can be used in linked data setting, and in user interfaces that can be built to access library data.
jindrich.mynarz@techlib.cz

Pardelli, Gabriella **145**
Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute since 1990, responsible for the Archives of the ILC Institute since 2005. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the Humanities. She has participated in many national and international projects including the recent projects:- BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council, (funded by CNR); - For digital edition of manuscripts of Ferdinand de Saussure (Research Programs of Relevant National Interest, PRIN - funded by the Ministry of Education, University and Research, MIUR). gabriella.pardelli@ilc.cnr.it

Pejšová, Petra **117**
Petra Pejsova studied information science and librarianship at Charles University. She works as an information specialist in the State technical Library, Czech Republic. Actually she is leading a project Digital Library for Grey Literature – Functional model and pilot. petra.pejsova@techlib.cz

Sassi, Manuela **145**
Manuela Sassi graduated in Foreign Languages and Literature at Pisa University, 110/110 cum laude. Since 1974 she has been working in Pisa at the Institute for Computational Linguistics of the National Research Council. Her interests and experiences range from linguistic to textual data processing and in providing linguistic resources on-line. She has been responsible for many national projects and has participated in numerous international projects.
manuela.sassi@ilc.cnr.it

Zoppi, Franco **149**
Franco Zoppi has been working for a long time in the design and implementation of software systems in the areas of DBMS, Distributed Office Information Systems and Digital Library Systems. Initially employed at Research and Development Department of Olivetti S.p.A., then at the Network Laboratory of the Telecommunications Department of Telecom Italia, in 2001 he joined the Information System Department of the University of Pisa as Project Manager. Since 2005 he has been working as Research Associate at the CNR-ISTI, where he coordinates the CNR activities in the BELIEF-II project. He is also involved in the DRIVER-II, EFG and HOPE FP7 projects. franco.zoppi@isti.cnr.it

https://doi.org/10.26069/greynet-2026-000.559-tgj

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-2026-000.559-tgj
TGJ Volume 6, Number 3 2010

An International Journal on Grey Literature

'RESEARCH ON GREY LITERATURE IN EUROPE'

AUTUMN 2010 – TGJ VOLUME 6, NUMBER 3