Repozitáře šedé literatury

Pejšová, Petra; Vaska, Marcus; Schöpfel, Joachim; Fürstová, Iveta; Polčák, Radim; Mach, Jan; Frantíková, Bohdana; Karlach, Petr; Dolanský, Jindřich
2010

Dostupný z `http://www.nusl.cz/ntk/nusl-97129`

Grey Literature Repositories

Editor Petra Pejšová

Authors:
Marcus Vaska
Joachim Schöpfel
Iveta Fürstová
Radim Polčák
Jan Mach
Bohdana Frantíková
Petr Karlach
Jindřich Dolanský

Prague 2010

National Technical Library
University of Economics, Prague

Publisher: Radim Bačuvčík - VeRBuM, Zlín 2010

CATALOGUING-IN-PUBLICATION – NATIONAL LIBRARY OF THE CZECH REPUBLIC

Grey literature repositories / editor Petra Pejšová;
authors Marcus Vaska ... [et al.]. - Zlín : VeRBuM, 2010 – 156 p.
Sub title: National Technical Library, University of Economics, Prague
ISBN 978-80-904273-6-5

08-027.571 * 004.08:930.25 * 005.591-027.542 * 024.5/.6:002.1 *
[024.5+659.2]-021.161 * 021.6 * 316.4.063.34 * 006.032/.033 * (437.3)
- grey literature
- digital archives – Czech Republic
- national programs and projects – Czech Republic -- 2001-2020
- accessibility of documents
- accessibility of information
- cooperative library systems
- international cooperation
- international standardization
- collective monographs

004 – Data processing, computer science [23].

The publication originated in the scope of the project *The Digital Library of Grey Literature - functional model and pilot realization*, supported by the Ministry of Culture of the Czech Republic in the framework of programme projects.

Reviewed by: RNDr. Miroslav Bartošek, CSc., Ing. Petr Žabička
The work was recommended for publication by the Scientific Editorial Board of VeRBuM publishing house.

First edition

ISBN 978-80-904273-6-5

Content

Preface

Grey literature, the topic of this publication, is generally considered an important source of up-to-date scientific findings by many authorities. At the same time, it is relatively difficult to locate and even more difficult to access. Thus, it may be supposed that all bodies funding research and development should eagerly support tools increasing the accessibility of grey literature - in order to reduce the probability of "reinventing the wheel" and to improve the efficiency of invested public assets. Five years ago, the EAGLE (European Association for Grey Literature Exploitation) collapsed - its landing lacked all eagle's majesty since it could not accomplish its main function, namely supply original documents. In response, we initiated a meeting of the representatives of the Research and Development Council of the Czech Republic, the Academy of Sciences of the Czech Republic, and university libraries. There, we presented a proposal on how to change the situation, at least at the national level. However, it is not enough simply to design and create a technical solution and demonstrate its usefulness to willing institutions. Five years have elapsed since that meeting and, by means of this book, I very much appreciate the long path that our project, the *National Repository of Grey Literature,* has taken since that time. I do not believe in the absolute efficiency of laws. However, any regulation which forces grey literature producers to make their materials available to the public may create a substantial amount of goodwill for the institutions. This has been clearly demonstrated regarding dissertations and other types of university final theses. While many countries make the results of publicly funded research freely available, in the Czech Republic, the only information available is that in official documents. A range of important types of grey literature has no legislative support and, thus, their accessibility is only dependent on the willingness of individual institutions and their awareness of the utility of cooperation among institutions. These are likely to increase, along with the increasing significance of this project. I deeply believe that the success of this project will serve as a strong argument for negotiating legislative support to reach the planned goals. Not for the fame of the *National Repository of Grey Literature*, but for better exploitation of the immense public funds invested into research and development. Hopefully, this book will also play its part in this task. I wish for it a multitude of kind readers.

Ing. Martin Svoboda

Director of the National Technical Library

Introduction

The potential of grey literature has recently attracted increasing attention by specialists working in the areas of education, research, development and information. This publication aims to enlighten the scientific community on grey literature and show a possible method for building a repository of grey literature. The scientific publication *Grey Literature Repositories ("Repozitáře šedé literatury")* is among the first publications of its kind in either the Czech Republic or abroad. It delivers comprehensive insight into building grey literature repositories in the Czech Republic. A similar scientific publication by Dominic J. Farace and Joachim Schöpfel called *Grey Literature in Library and Information Studies* was issued abroad in September 2010.

The book *Grey Literature Repositories* provides information about the concept of grey literature, its development, types, and producers in the European context, with a focus on the Czech Republic. The example of the National Repository of Grey Literature, practically implemented in the National Technical Library, is given to illustrate how such a repository may be built. The most important factors in the success of such project are cooperation and legislation. The book presents the components necessary to construct a repository from a practical viewpoint, namely analysis, selection and implementation of software, creation of the metadata format and audit of repository credibility.

Chapter 1 introduces the topic of grey literature, its history, types, and development. It is also devoted to grey literature producers, who are rather diverse compared to the producers of classical "white" literature. In the second section, features of grey literature and its impact on the scientific public and public at large are analyzed. Examples of grey literature in various scientific areas are given in conclusion.

Chapter 2 describes the state of grey literature in Europe. Examples of mandatory policies are stated, as well as strategies for acquiring grey literature. Outstanding grey literature collections, both printed and electronic, are presented. This chapter highlights the importance of cooperation for the acquisition, storage and accessibility of grey literature at all levels, from departmental to national and international collections.

The development of grey literature is summarized in Chapter 3, along with current systems and projects involving cooperation. It depicts the role both historically and presently of the National Technical Library in supporting the acquisition of grey literature in the Czech Republic and participation in international cooperation systems. As part of these activities, the project of the National Repository of Grey Literature has been underway since 2008.

Cooperation with producers is a key activity for repository construction. Although grey literature is a precious source of information, it is rather difficult to capture using classical means of collection. Contacts, partnership and networking are of immense importance. Moreover, cooperation may prove useful in sharing the development of standards, methodologies and tools. This is because these activities are so demanding both for funding and capacity that they cannot be reasonably pursued at each separate institution. Importantly, every area of grey literature should be supervised by one particular organization with support from others.

University final theses are such a distinctive type of grey literature, from a legal standpoint as well, that an entire chapter is devoted to them. Access to final theses varies widely from university to university. University representatives have taken advantage of international experience to create the EVSKP-MS metadata format. Cooperation between universities resulted in construction of the *National Registry of Theses* in 2008.

The legal view of grey literature is of crucial significance. This book is written with a view to Czech law, which has specific characteristics compared to the laws of other countries. From a legal standpoint, rather than an exact definition of grey literature, it is important to specify which legal classification applies to particular species of grey literature. This chapter explains basic expressions in this area, such as authors' rights, legal licensing, free licenses, copyright, academic work, employee work, and exclusive licenses. In conclusion, legal recommendations for managers of grey literature databases and options for exploiting free licenses are debated.

Chapter 7 is entirely given over to descriptive elements, so-called metadata. The metadata format created for the *National Repository of Grey Literature* is described, its division into obligatory and specific elements, as well as technical and administrative metadata. In addition to the description of the metadata format, this chapter provides an overview of the basic model metadata formats, i.e., Dublin Core, OpenSIGLE, ETD-MS, MARC21, EVSKP-MS, and ASEP.

The selection of a software system is a key step in constructing a repository. The software requirements for the *National Repository of Grey Literature* were divided into thirteen groups and may also be used to select institutional repositories. Since the software requirements for digital repositories are very well covered by OpenSource applications, the analysis of the most common systems is presented, these being CDS Invenio, DSpace, EPrints, Fedora, and Greenstone.

The software solution developed for the pilot implementation of the *National Repository of Grey Literature* is composed of two parts. The first part is represented by CDS Invenio for the digital repository; the other part is formed by the ESP FAST indexing and search system. The preconfigured local

installation of CDS Invenio for grey literature collections is offered to partner organizations to be installed over their own hardware. The major portion of the chapter is devoted to this system. The *National Repository of Grey Literature*, as well as other repositories of scientific grey literature in the Czech Republic, may be searched using a central search interface run by the ESP FAST system.

The final chapter describes the audit of the *National Repository of Grey Literature* on the issue of its trustworthiness. This audit was run internally with the help of the DRAMBORA tool. The chapter briefly summarizes the audit procedure and the resulting list of risks identified. As part of the audit, strategies are defined concerning how to manage risks, prevent them, and potentially resolve their consequences. The DRAMBORA tool is available online and enables continuous updating and evaluation of the audit.

On behalf of our team of authors, I wish you interesting reading.

Petra Pejšová

1. Introducing Grey Literature

Marcus Vaska

A few years ago, reference to the concept of "grey literature" would perhaps only have been fully appreciated by a tight, close-knit community of researchers. Some may see this concept as referring to sadness, difficult times, and even questionable authority. There are also those individuals that view this body of literature as purely scientific, referring to the brain's grey matter and thus encompassing knowledge that is highly intellectual (Mason, 2010).

Today however, whether one hears graue Literatur, Letteratura grigia, literature grise or šedá literatura, one can easily recognize what these words are referring to. It is therefore remarkable how efforts over the past two decades have altered originally preconceived notions of grey literature, resulting in a deeper understanding of how the formation, retrieval, and use of non-traditional material becomes a fundamental companion in one's research endeavors. The advent and use of new technologies, a prevalent trend in 21st century society, has created increasing awareness of the need to incorporate the grey alongside the white, a boundary that is becoming less distinct and which may become translucent in the not-too-distant future.

The Concept of Grey Literature: A Definition that is truly "grey"

The task of creating a succinct definition of grey literature has been just as challenging as the discussion of how to raise awareness of the existence of this wide body of publications. Some researchers are euphoric when discussing the use of grey literature in their own research, claiming that it is "the unsung hero, the foot soldier, the foundation of the building" (Mason, 2010). While such an all-praising statement undoubtedly gives support to the importance of this material, low print-runs and small distribution patterns of grey literature can cause a tendency to overlook the role of this material in the search process.

The rise and uncovering of grey literature has been met with numerous attempts on how best to explain and name this concept. This type of literature has so often been simply called "grey" that numerous other synonyms for this term may be overlooked when in fact they exist. These include non-conventional, informal, informally published, fugitive, and invisible. Auger (1994, p.3) elaborates on this further by stating that "documents may be unconventional in many ways, and that many conventionally published documents show greyish aspects".

Several definitions of grey literature have been put forth, indicating that, by its nature, this is a body of literature that is indeed difficult to define. Weintraub (2000) sees grey literature as "publications issued by government, academia, business, and industry, in both print and electronic formats, but not controlled by commercial publishing interests, and where papers, theses, government documents, bulletins, fact sheets, conference proceedings and other publications (are) distributed (for) free, available by subscription, or for sale." In comparison, a posting on Wikipedia suggests that grey literature "is a term used variably by the intelligence community, librarians, and medical and research professionals to refer to a body of materials that cannot be found easily through conventional channels such as publishers" (Gray Literature, 2010). Nevertheless, an explanation put forth more than a decade ago by the Third International Conference on Grey Literature (abbrev. GL'97) is perhaps the one that has been mostly accepted in academia: "that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers" (GreyNet, 1997).

The above-mentioned characterization emphasizes that researchers should include grey literature and published material side-by-side whenever conducting literature searches (especially when performing systematic reviews), in order to limit bias. The researcher should always strive to remain objective in the undertaking of an information-seeking pursuit; failure to heed various forms and formats of literature can lead to a subjective one-sided research path.

A common misconception that some share towards the field of grey literature is that it is literature that is only partially published, since it is produced on a relatively small scale, and is not as widely available as material readily accessible in the main streaming publishing world. However, according to the origins of the word, to publish is to make publically available; therefore, everything that one produces is in effect a published work, regardless of whether or not one is aware of its existence.

Grey Literature: History and Development

Depending on how broadly one wishes to apply the previously mentioned grey literature definition, it can be argued that theories speculating the existence of this literature form since Roman times certainly do hold merit. In fact, some would consider the notes of Archimedes, Socrates, and other like-minded ancient scientists and philosophers to have already exposed the world to a crude form of grey literature (Auger, 1989). Grey literature is therefore not a new concept, although these types of documents were not always labeled as "grey". In fact, grey literature used to be referred to as Reports Literature.

One of the earliest grey literature reports, *The Behaviour of Aeroplanes in Gusts*, by NASA, arose from the aviation industry (Auger, 1989, p.13). The first recorded technical report of grey literature took place in Great Britain military activities in 1909 (*Research Memoranda of the Aeronautics Research Council),* as well as in 1915 by the U.S. National Advisory Committee for Aeronautics. WWII caused a prolific rise in Reports Literature; this documentation served as a major means of communicating research-testing results. During these times of war, there was a greater need for security and confidentiality; documents had to be produced secretly, not in the conventional publishing manner (Auger, 1989). As Gelfand (2000, p. 74) further points out, "the roots of grey literature lie deep in the intelligence and government research communities", and will, in the majority of cases, continue along this path.

A few decades ago, grey literature publications were seen as separate entities from the mainstream commercially available books and journals; they were not considered to be permanent parts of a collection. Yet, if one considers the increasing number of records found in library catalogues that cite material of a non-traditional nature, this notion becomes increasingly difficult to fathom. Grey literature has undergone leaps and bounds from its recognition as a vital aspect for many organizations in the 1970s, to its official acceptance as a necessary component of effective literature searching in the 1990s (Auger, 1989).

Types of Grey Literature

Considering the technology-enhanced society in which we live, grey literature has taken on many forms. Regardless of the format within which these documents may appear, those that make regular use of grey literature material "consider [it] to be a primary source of information" (Alberani, 1990, p.358). Aside from the traditional print publications such as theses and dissertations, conference papers and proceedings (both published and unpublished versions), translations (scientific and technical literature), newsletters, and government reports, electronic means of accessing this material have risen at an alarming rate. Statistics, data sources, bulletins, course materials, company literature, library catalogues, digital libraries, patent documents, and conference proceedings are, for the most part, available and accessible on the World Wide Web, due in large part to the power and comprehensiveness of Google and other federated search engines. The advent of new formats of distributing this information, namely multimedia, social networking, and various forms of informal, yet informational communication, including telephone calls, meeting minutes, e-mail, interviews, and even postings on the Wikipedia, the collaborative encyclopedia, have created new possibilities to explore.

Another type of grey literature that is gaining wider acceptance is the production and distribution of pamphlets, posters, tickets, and timetables. Auger (1994, p. 7) refers to this as ephemera, "material which carries a verbal message and is produced by printing or illustrative processes, but not in the standard book, periodical, or pamphlet format."

Undoubtedly, the rise of social networking tools as reliable sources of information has caused considerable debate in the scientific community. Proceeding with caution and scrutinizing the trustworthiness of the material presented on sites such as Twitter, Second Life, and Facebook have caused some in the field of academia to discount these sources as representative of grey literature. However, the function of these multimedia outlets is to provide timely current awareness material, a feature of grey literature that cannot be overlooked. H1N1, a worldwide pandemic that broke out last year and which threatens to reemerge once again, had a very real presence on Twitter, particularly during the early months when healthcare officials were scrambling to organize a mass vaccination program. In fact, a number of the tweets were actually quite informative, and were accessible to the public long before an article was published in the daily newspaper.

Despite the power of the digital age, there is still a considerable amount of grey material that has not yet transcended the boundary and become retrievable in an electronic format. The age-old paradigms of hand-searching through journals, browsing through a bibliography, and speaking directly to researchers are all valid traits that are bound to continue for some time to come (Helmer, 1999).

Grey Literature Producers

The plethora of types of grey literature in existence today gives rise to an ever-increasing multitude of organizations responsible for creating, delivering, and maintaining grey literature content. From private publishers and societies to associations, councils, research establishments, federations, institutions, trade unions, trusts, and universities, each content creator has played a pivotal role in raising grey literature awareness.

The Open Access movement, whereby content to scholarly material is freely available on the World Wide Web, constituting free access for the user, has resulted in an influx of grey literature digital libraries and information repositories. Whether these producers are known only on a smaller local scale, such as the *Grey Literature LibGuide* developed at the *University of Calgary*, reach a wider national audience, namely the *Association of Libraries of Czech Universities* (abbrev. AKVŠ) or the *Academy of Sciences of the Czech Republic* (abbrev. AS CR), or achieve worldwide recognition as prominent national and international repositories, such as the Grey Literature Network

Service[1] (abbrev. GreyNet), or OpenSigle, a system for information about grey literature in Europe, the goal remains fundamentally the same: "to facilitate communication between individuals and organizations producing grey literature" (International Organizations-NUŠL, 2009). The list of grey literature databases in the Czech Republic and abroad is summarized in Appendix 1 and 2.

Features and Impact of Grey Literature

The features of grey literature can be broken down into the following key traits: rapid publication, variable formats (flexibility), detailed, no peer review, no commercial source, and, in most cases, limited distribution. Speed and flexibility provide a distinct advantage for grey literature material, as it allows for specific, focused discussion on perhaps less well-known topics to be produced. As has been indicated by the multitude and diverse types of grey literature, this literature forms a part of one's daily life (from the newspaper that is read in the morning, to e-mail and Internet activities that the vast majority of society takes part in on a daily basis). In fact, Coonin (2003) boldly proclaims that "everything we read outside of journals and books can be considered grey literature." With this idea firmly entrenched in one's mind, it becomes rather easy to see why the emergence and growth of grey literature has been "three to four times that of conventional literature" (Farace, 1997). Over a span of nearly two decades, GL conferences have been held around the world, discussing the impact of grey literature on the information society.

Instantaneously accessing information is useful in both the physical and virtual world. Despite the explosion of grey literature due in large part to numerous technological enhancements (namely the Internet), some critics remain skeptical as to how informative literature that is not peer-reviewed, is often scrutinized, and does not imply a qualitative nature, truly is. Dominic Farace, a renowned leader in the grey literature field, argues that this mentality is counter-productive to grey literature's aim to create greater recognition among its users: "just because grey literature is not controlled by commercial publishers does not mean that it has not undergone as severe a peer process as does commercially published materials" (Gelfand, 2000, p.74). Perhaps Outten (2008) puts some closure to this debate in her arguments that although not considered to be a scholarly form of publication, grey literature is produced by experts in the field, and "serves scholars and lay readers alike with research summaries, facts, statistics, and other data that offer a more comprehensive view of the topic of interest." With the growing interest in grey literature being demonstrated in a number of areas, including distance education, grey literature can simply not be overlooked.

[1] http://www.greynet.org/

However, the identification of relevant studies in grey literature and their inclusion in systematic reviews can be particularly time-consuming and difficult. There is also some controversy as to whether unpublished studies should be included because they might be incomplete and their methodological quality can be difficult to assess (Helmer, 1999). This therefore leads to the following dilemma: with easier access to grey literature documents than previously possible, has the quality been maintained? Due to its interdisciplinary nature, grey literature can comprise more than one term for a particular document. Studies have shown that search engines typically index only 16% of all available web content; nevertheless, these electronic tools are what users turn to first in an attempt to retrieve elusive items (Helmer, 1999). If information is not adequately indexed, it becomes virtually impossible to find. This challenge is a direct correlate of the need to maintain repositories of grey literature material, such as GreyNet, in order to address and combat the Invisible Deep Web access issue. This methodology showcases the significance of utilizing resources other than online databases to add greater comprehensiveness to a search; studies have indicated that "an additional 29.3% of items were uncovered by using extended search methods in addition to the mainstream sources which were searched" (Helmer, 1999).

Understandably, with the billions of pages circulating on the Internet, the content and/or quality cannot always be easily verified. Researchers using these means of information retrieval should be cautious and not take everything they find and read at face value. However, one fundamental feature of grey literature material residing on the Web is that this material achieves a high level of rapid publication, often being posted as soon as it is written.

Above all, lack of bibliographic control, non-professional format, and limited print runs serve as key issues of concern in the field of grey literature. (Auger, 1989) With a less than obvious title and poor descriptive information, Auger further goes on to say that many publications use the minimalist approach of indexing, providing "few details, no abstracts, and no indexes" (p.6).

With the potential of such a wide berth of information available at one's fingertips, determining where to begin a search can be a daunting task. While browsing through an organization's website and scanning through bibliographies and reference lists may be a starting point, not all documents are available online (some are still produced only in print). Therefore, it may be necessary to hand- search relevant journals, especially if they are not indexed in a primary database vendor. Finally, a last resort may involve contacting the author or organization of a particular publication directly. In addition, for many systematic reviews, dissertation abstracts may well be the starting point, before embarking on a search for additional grey literature on a topic.

Grey Literature in Various Disciplines: Building Information Repositories

Despite the wealth of grey literature material, creating awareness and access to this material is an obstacle that is yet to be completely overcome. The notion of open repositories, whereby information is freely available on the Web, is a phenomenon that is gaining widespread acceptance (Shearer, 2010). Shearer (2010, p. 25) points out that building repositories of grey literature is possible "because the majority of journals (about 70%) permit authors to archive copies of their articles in a repository". The following examples provide a brief synopsis of the importance and impact that open repositories play in different subject specialties in the grey literature neighborhood.

"Published trials tend to be larger and show an overall greater treatment effect than grey trials" (Hopewell, 2007). Therefore, reviewers must ensure that both published and unpublished trials are evaluated objectively without any bias. This can often be a challenging task, especially in the healthcare sector, where daily decisions directly affect patient lives. Further, failure to consider grey literature data in assessing patient needs may result in over-confident estimates of treatment effectiveness, particularly if results are solely based on trials published in mainstream journal publications. Rather, it becomes necessary to carefully and thoroughly explore government resources (local, national, and international), reports, registries, associations, clinical trials or guidelines, as well as databases that provide evidence-based recommendations. While these publications may not be as well known as their commercially published counterparts, alternate views of treatment effectiveness that have tendencies to be overlooked need to be included in order to ensure comprehensiveness. Thus, issues of granting and/or restricting access to electronic medical records, itself a form of grey literature, is playing a substantial role in the realm of health technology assessment (Helmer, 1999).

The transient nature of posters, programs, and other such publicity materials make grey literature a key component of material within the performing arts. These documents are often the only means that users can refer to when searching for particular playwrights, performers, or artists (Grey Literature Speaker Series, 2009). In contrast, the legal world must deal with court documents, committee reports, policy documents, working paper series, and of course, case studies on a daily basis, all of which are types of grey literature.

While the above examples have illustrated the importance of grey literature in a particular subject domain, the interdisciplinary nature of research pursuits must also be considered. The conjoined efforts of humanities and social science divisions have caused the emergence of several documents

depicting the atrocities of the Holocaust, putting "a face on what was lost, allowing readers and researchers to grasp the enormity of what happened to six million Jews" (Grey Literature Speaker Series, 2009). The eyewitness accounts, a repository of grey literature material in its own right, makes sure that this period of racial persecution is never forgotten.

Maintaining a Stronghold on Grey Literature: Future Possibilities

While it is refreshing to witness more and more students, faculty, and researchers take interest in the world of grey, the task of raising grey literature awareness is far from complete. Despite our most valiant efforts, there is still a multitude of grey literature documents that exist only on paper; much of this material is still not available online. Nevertheless, new advances in technology are enabling grey literature researchers to come forward and share their expertise with the wider community. As the theme of GL12 and the compilation of this monograph entails, grey literature is indeed a collaborative venture, "maintained by a community of scholars, with assistance from a university library or technical services" (Shearer, 2010, p.28).

A movement that began during the final decade of the 20th century, the prevalence of grey literature in the information society is undeniable. As the pondering over the impact of grey literature continues, it is perhaps fitting to end with Farace's vision of grey literature that he presented more than 10 years ago: "just as in the last decade of the twentieth century came to dominate the information supply side, so well before the end of the first decade of this twenty-first century society will also come to dominate the information demand side" (Gelfand, 2000, p.76).

REFERENCES:

ALBERANI, V. - de CASTRO PIETRANGELI, P. - MAZZA, A. *The use of grey literature in health sciences: A preliminary survey.* Bulletin of the Medical Library Association, 1990, vol. 78, no. 4, p. 358-363.

AUGER, C. *Information sources in grey literature* (2nd ed.) London: Bowker-Saur. 1989.

AUGER, C. *Information sources in grey literature* (3rd ed.) London: Bowker-Saur. 1994.

CARROLL, B. - COTTER, G. *A new generation of grey literature: The impact of advanced information technologies.* Publishing Research Quarterly, 1997, vol. 13, no.2, p. 5-14.

FARACE, D. *Rise of the phoenix: A review of new forms and exploitations of grey literature.* Publishing Research Quarterly, 1997, vol. 13, no. 2, p. 69.

GELFAND, J. *Interview with Dominic Farace, founder of GreyNet*. The International Journal on Grey Literature, 2000, vol. 1, no. 2, p. 73-76.

GreyNet. *Perspectives on the Design and Transfer of Scientific and Technical Information*. GL'97: Third International Conference on Grey Literature, Luxembourg 1997.

HELMER, D. - SAVOIE, I. - GREEN, C. (British Columbia Office of Health Technology Assessment). *How do various fugitive literature searching methods impact the comprehensiveness of literature uncovered for systematic review?* Vancouver: British Columbia Office of Health Technology Assessment. 1999.

HOPEWELL, S. - McDONALD, S. - CLARKE, M. - EGGER, M. *Grey literature in meta-analyses of randomized trials of health care interventions*. Cochrane Database of Systematic Reviews, (2). 2007.

SHEARER, K. *A review of emerging models in Canadian academic publishing*. University of British Columbia: University of British Columbia Library. 2010.

ELECTRONIC RESOURCES:

COONIN, B. *Grey literature: An annotated bibliography* [online]. 2003. Retrieved March 20, 2010 from http://personal.ecu.edu/cooninb/Greyliterature.htm.

Grey Literature [online]. 2010. Retrieved March 20, 2010. Available from http://en.wikipedia.org/wiki/Grey_literature.

Grey Literature Speaker Series [online]. 2009. The University of Calgary. Retrieved March 20, 2010 from http://glspeakers.wetpaint.com.

International organizations - NUSL [online]. National Technical Library. 2009. Retrieved June 10, 2010 from www: http://nrgl.techlib.cz/index.php/Mezin%C3%A1rodn%C3%AD_organizace.

MASON, M. *Grey literature: History, definition, acquisition, and cataloguing* [online]. 2010. Retrieved March 20, 2010 from http://www.moyak.com/papers/grey-technical-literature.html.

OUTTEN, C. *Grey literature* [online]. 2008. Retrieved March 20, 2010 from http://www.csulb.edu/library/subj/gray_literature.

2. Access to European Grey Literature

Joachim Schöpfel

Networking

Collecting grey literature remains a challenge to library and information science (abbrev. LIS) professionals. Grey items such as reports, proceedings, or working papers cannot be purchased or bought like journals and books. There is no special agency or supplier for grey materials.

Buying information is part of the traditional library role, together with gateway and archive functions. In line with the economic definition of grey literature, "material that usually is available through specialized channels and may not enter normal channels (…) of (…) distribution" (Schöpfel & Farace, 2010), one comes to understand that a systematic collection of grey literature calls on specific attention, competency, and procedures.

Frequently, grey holdings are the result of a patient and long-term investment in professional contacts and networking. Networking means sharing information about grey content, resource discovery and acquisition channels with other LIS professionals. Two recent personal Web 2.0 initiatives of sharing insider knowledge on biomedical grey literature are Barrera's personal Netvibes page with *Really Simple Syndication* (abbrev. RSS) files[2] and Giustini's report *Finding the Hard to Finds* on grey holdings and search strategies.[3]

Yet, LIS professionals started collaborative "grey work" many years before Tim O'Reilly invented the Web 2.0. In 1980, EU scientific information centres established the *System for Information on Grey Literature in Europe* (abbrev. SIGLE) to provide access to European grey literature and to improve bibliographic coverage. After initial funding by the *Commission of the European Communities*, the national centres formed a non-profit network for the acquisition, identification and dissemination of grey literature called *European Association for Grey Literature Exploitation* (abbrev. EAGLE). In this network, each national centre for scientific and technical information (STI) held the national grey collection or at least, guaranteed the document supply of distributed holdings.

In 2005, a general assembly resolved to liquidate EAGLE because its organisational structure and business model were unable to cope with Internet technology and the Google generation, e.g. SIGLE offered no

[2] http://www.netvibes.com/ernestobarrera#Grey_literature

[3] http://www.slideshare.net/giustinid/finding-the-hard-to-finds-searching-for-grey-gray-literature-2010

solution for online cataloguing, metadata harvesting, links to full-text and other resources (Schöpfel et al., 2007).

But the same 2005 EAGLE general assembly decided unanimously to preserve the European co-operation for grey literature and to transform the 1980 model into a sustainable network in the emerging environment of open access to scientific information, especially in the context of the 2003 Berlin Declaration.

The first step was to archive the SIGLE records in an open and freely searchable database, compliant with the OAI metadata harvesting protocol.[4] The French *Institute for Scientific and Technical Information of the French National Center for Scientific Research*[5] (abbrev. INIST) developed OpenSIGLE[6] based on *Massachusetts Institute of Technology* (abbrev. MIT) software (DSpace) and loaded most of the SIGLE records in a simplified XML format (Farace et al., 2009). The second (future) step should be the federation of European open access projects for grey literature in order to (re)establish a gateway to European grey literature.

Other grey networks may exist on a regional or community level. For instance, in the early 90s the French government operated an interdepartmental group *Littérature Grise Administrative* (abbrev. LIGRIA) for the management of official grey literature.

What can be learnt from these and other initiatives? Most LIS professionals in charge of grey collections are interested in collaborative work. Librarians like networking. Cultural mediation is collective, not solitary. But "network attitude" is not enough. Efficient networking needs experience and competency, common objectives, well-organized governance structures and a sustainable business model. This may explain why networking sometimes remains an individual (personal) rather than an institutional affair.

Mandatory policies

Some collections of grey literature are the expression of a clear mandatory policy, e.g. result from explicit national or regional decisions. These may be scientific, cultural and/or political decisions, for instance to guarantee preservation of and access to specific contents, or to contribute to the construction of cultural (scientific) heritage.

The first case is a kind of legal deposit of grey items. One of the three special scientific libraries in Germany, the *German National Library of Science and Technology*[7] in Hanover (abbrev. TIB) celebrated its 50th

[4] Open Archives Initiative Protocol for Metadata Harvesting
[5] http://international.inist.fr/
[6] http://opensigle.inist.fr
[7] http://www.tib-hannover.de

anniversary in June 2009. TIB defines itself as a transfer centre for scientific knowledge; its task is "to comprehensively acquire and archive literature from around the world pertaining to engineering and the natural sciences".[8] The library places a particular emphasis on the acquisition of grey literature (conference proceedings, research reports, standards and dissertations in print and digital format). Its grey holdings are unique in Germany. In 2010, TIB holds more than 210.000 print and 30.000 digital German research reports on engineering or natural sciences. Each month, around 200 new electronic and 500 print reports are added. TIB is the deposit library of the (digital) final project reports funded by the *Federal Ministry of Education and Research* (abbrev. BMBF[9]). Since 1996, any research project has to deliver its final report to the TIB both as a free, printed copy, and on an electronic storage medium (see also Meyer, 2009). – The *Moscow Scientific and Technical Information Centre* (abbrev. VNTIC) receives copies of Russian PhD theses; it already holds 500,000 theses since 1982.

A quite different model of mandatory policy is the distributed collection of Ph.D. theses by academic libraries, with a central access point. The French government published in 1985 a decree that regulated and improved the deposit and dissemination of doctoral theses. The local library stores the document but the record is part of the French national union catalogue *Système Universitaire de Documentation*[10] (abbrev. SUDOC) that allows for ordering and delivering of print copies (Paillassard et al., 2007).

A third model are bi- or multilateral agreements for the acquisition and dissemination of grey items in the context of a national STI policy. Again, a French example may illustrate this model: one part of the significant INIST holdings of French Ph.D. theses and scientific reports are related to settlements with the Ministry of Higher Education (theses) or publishing bodies (research organisations, ministries). Another example: the Danish Royal Library was a depository library until July 2002 for the Council of Europe, United Nations, NATO, OECD, UNESCO and other international organisations.

Especially when accompanied by public funding, these explicit mandates allows for long-term collection and preservation of specific grey items. Sometimes they will also facilitate digitization projects (scientific heritage). The problem with these public mandates and agreements is that they may be under-funded and/or of limited duration, with a risk of incomplete and disrupted holdings.

[8] http://www.tib-hannover.de/en/the-tib/overview-of-the-library/ [consulted 26 August 2010]

[9] Bundesministerium für Bildung und Forschung

[10] http://www.sudoc.abes.fr

Acquisition strategy

Defining a coherent acquisition policy is a crucial part of a library's function. Often, this policy will reflect patrons' needs and suggestions, subject choices and budget structure. Some libraries may also develop a specific acquisition strategy with regards to grey literature, independently from an external mandate. Over the time, such an institutional ("intrinsic") strategy may generate exceptional holdings.

Probably one of the most famous holdings of this kind is the *Boston Spa conferences collection* (abbrev. Boston Spa) of the British Library with around 450,000 items. "British Library holds one of the most comprehensive and easily accessible collections of English language conferences in the world. (…) The British Library believes holding the material is only part of the process to enable access, and has developed various products to aid the user in locating this material" (Tillett & Newbold, 2006). This holding reveals an explicit choice to collect all English-speaking scientific conference proceedings. During the golden age of document supply, LIS professionals and customers knew that Boston Spa possessed (nearly) all international scientific conferences.

Another distinct area of the collecting and focus maintained at Boston Spa are scientific and technical reports, from several thousands public and private sector British, American and international sources.[11] The 2008-2009 British Library annual report mentions 10,5 million reports on microform while the website provides the figure of 4,9 million unrestricted reports available for public use.

As a complement to their mandate for German reports, the TIB invests in the systematic collection of foreign scientific reports, especially by the *National Technical Information Service* (abbrev. NTIS) of the US Department of Commerce or the NASA but also from an important number of European institutions (nearly 2 millions).

These two examples show, too, that up to now we cannot speak of a formal European coordination of national or regional grey collections. National or institutional considerations prevail, e.g. preservation strategies, national independency, bilateral agreements etc.

Somewhere on the crossroad between mandate and institutional choice we find holdings of theses and dissertations. Usually, academic libraries are mandated to hold theses from their own university; at the same time, they collect more or less systematically theses from other universities, following their own rules and criteria (disciplines, subjects, institutions etc.). At first sight, this does not make any difference. But when libraries evaluate and

[11] NASA, NTIS, ERIC, DOE, FAO, INIS, ESA etc.

weed their collections, they will maintain the "local collection" and try to discard the rest.

From print to digital collections

Since the invention and success of the web, libraries leave the Gutenberg galaxy with its millions of print items. They do it in two ways. They convert their print holdings into digital collections, and they collect and archive born digital material.

Grey literature enters the two circuits. Significant retro-digitization projects were launched for PhD theses in print format or on microforms. The British Library digitizes theses from UK universities for the new EthOS-portal[12] that integrates free access to electronic theses and dissertations (abbrev. ETDs) harvested from open repositories with supply from theses digitized on demand.

In France, the national reproduction centre for theses *L'Association Nationale de la Recherche et de la Technologie*[13] (abbrev. ANRT) develops a capacity for digitisation for its service *Thèses à la carte*. The German digitization projects funded by the *Deutche Forschungsgemainschaft*[14] (abbrev. DFG) contain namely early primary sources (cultural heritage) but also, scientific publications (manuscripts, journals etc.). In the UK, the *Joint Information Systems Committee* (abbrev. JISC) invests since 2003 in digitizing content from special collections[15] like the approximately 600 volumes of historical population reports (census reports) hold by the *University of Essex* and 10,000 theses for EthOS.

The list of European digitization programs for grey literature is long, and we could add the retro-digitization programs by the *Delft University of Technology* (abbrev. TU Delft) for their E-thesis pilot, by the University of Uppsala for more than 11,000 theses submitted in the 18th and 19th century, the Catalan electronic theses and dissertations network *Tesis Doctorals en Xarxa* (abbrev. TDX), or the Poznan PSNC[16] *Digital Libraries Team* project coordination activity.

In comparison, less digitization programs are scheduled for reports, conference proceedings or other forms of grey literature. Also, we didn't hear from important European initiatives that could match with the US digitization project of *Department of Energy* (abbrev. DOE) report collections or the *Office of Scientific a Technical Information* (abbrev. OSTI)

[12] http://ethos.bl.uk/
[13] http://www.anrtheses.com.fr/
[14] http://gepris.dfg.de
[15] JISC digitisation and e-content program, http://www.jisc.ac.uk/digitisation.
[16] Poznan Supercomputing and Networking Center, http://www.geysers.eu/index.php/theproject/partners/psnc.

collaboration with other sites (FERMI, LANL etc.).[17] Three recent initiatives in France involve the *Bureau de recherches géologiques et minières* (abbrev. BRGM) reports for an earth sciences portal, the LARA[18] platform for scientific reports from different institutions and the mathematics archive *Numérisation de documents anciens mathématiques*[19] (abbrev. NUMDAM) with 29 seminaries from 1948 to 2007.

These programs share three common features, e.g. clearly identified grey collections, scientific heritage character and low coordination with other initiatives.

Sometimes, digitized grey items are mingled up with born digital material and/or living e-collections. One (but not the only) example is the French national repository for ETDs, *Theses en Ligne* (abbrev. TEL), with more than 10,000 recent theses (2005-2010) but nearly 2,000 digitized PhD theses published in 1990 or before.

Another site, the UK *Centre for Environmental Data Archival*[20] (abbrev. CEDA) based at the STFC[21] *Rutherford Appleton Laboratory*, is for grey literature primarily concerning Earth observation and the atmospheric sciences. Apparently, all CEDA items – more than 600 - are born digital. Other European repositories with grey digitized and/or born digital material can be found in the *The Directory of Open Access Repository* (abbrev. OpenDOAR) directory: from the 776 registered sites (March 2010), 54% hold ETDs, 42% unpublished reports and working papers, and 40% conference and workshop papers. In France, ¾ of the OA repositories contain grey literature.

The open access principle

The OpenDOAR directory highlights the fundamental impact the shift from print to digital holdings produced on the underlying business model and the distribution channels of grey items.

Grey literature is defined through its non-commercial dissemination channels. With the development of the open archive (OA) initiative, grey documents quite logically took their place in these new repositories, especially in institutional repositories (Schöpfel et al., 2009) but also in subject-based or other types of open archives.

[17] See for instance Fermi's 1947 report on the Future of Atomic Energy available at http://www.osti.gov/accomplishments/documents/fullText/ACC0043.pdf.
[18] Libre accès aux rapports scientifiques et techniques, http://lara.inist.fr/ see also Stock et al. (2006)
[19] http://www.numdam.org/
[20] http://cedadocs.badc.rl.ac.uk/
[21] Scientific a Technology Facilities Council, http://www.scitech.ac.uk/Home.aspx.

A longitudinal survey 2005-2009 describes how five international STI centres adopt a strategy of open access publishing, in different environments, with different objectives, and with more or less success (Boukacem-Zeghmouri et al., 2006; Schöpfel et al., 2009). The total number of items freely available through their open repositories is difficult to estimate; it may be approximately 3.5 to 4 million items including a significant part of grey items. This special material is by definition part of the long tail – a lot of items with a low demand. Luzi et al. (2008) describe the preparation of an institutional repository by the Italian *National Research Council* (CNR);[22] at least one third of the deposits in existing CNR OA sites are grey items (reports, oral presentation, theses etc.).

The open archive may be the best solution for this kind of "stuff" because of limited acquisition, management, conservation and supply costs. Yet, this remains an assumption without empirical support because there is no economic or financial evidence so far as we know.

In the next future, will all grey documents be available on OA web sites? In spite of the Willinsky (2006) claim that "open access is a public good" and that "commitment to scholarly work carries with it a responsibility to circulate that work as widely as possible", a significant part of grey material probably may not enter the open archive landscape – because of lacking interest or budget for digitization of older print materials or restricted access, or because the items are already available on other websites (personal pages, institutional sites with links to PDFs etc.) but not necessarily well indexed.

Nevertheless, the proportion of "grey" documents published on the Web continues to increase. This development is closely linked to the production of grey literature in e-environments, as well as to retrospective activities leading to republication. The Internet will encourage a greater diversity in the types of "grey" resources available (raw research results, notes and personal comments, lectures, newsletters, product catalogs etc.).

New technologies of information and communication facilitate resource archiving in general, and there is strong incentives from the OA movement. Nevertheless, the question of "who should archive what, where, when, and for how long" remains largely unanswered. Aware of information policy and the concomitant financial aspects involved, answers are rather urgently needed, even if they were to now address only part of grey literature resources (Schöpfel et al., 2010).

[22] Consiglio Nazionale delle Ricerche

From collection to gateway

"A library is a collection of sources, resources, and services (…); it is organized for use and maintained (…)."[23] Can an open repository be called a collection? Is it part of the library collections?

Probably, these are yesterday questions. In the coming Google era, do we really need well-structured and maintained grey collections? Or do we need tools to search, retrieve and access grey items? Can we imagine grey "collections" as a kind of global grid?

Instead of answering, we would like to draw the reader's attention to some recent developments, products and services.

A couple of years ago, the *Royal Netherlands Academy of Arts and Sciences* (abbrev. KNAW) stopped all acquisition activity of the *Institute for Scientific Information Services* (abbrev. NIWI), formerly one of the major document suppliers. Instead, the KNAW invested in the creation of a new gateway called *National Academic Research and Collaborations Information System* (abbrev. NARCIS) that gives access to OA publications from Dutch universities and research institutes, datasets, descriptions of research projects, institutes and researchers, and research news. In this environment, the borderline between "grey" and "white" (commercial) literature becomes increasingly indistinct.

The *National Documentation Centre*[24] (abbrev. EKT) at Athens maintains the *Hellenic Dissertations*[25] database linking to 13,000 theses hold by Greek universities.

The *Eidgenössische Technische Hochschule* (abbrev. ETH) Zurich library portal provides access to 2,1 million reports from other libraries, databases and search engines.

The *Irish Virtual Research Library and Archive* is meant to realize the latent potential of archival collections within the *University College of Dublin*.

The *Academic Archive On-line*[26] (abbrev. DiVA portal) gives access to 270,000 research publications and student theses written at 27 Swedish and Norwegian universities and colleges of higher education; 44% of the content is grey.

Scirus, *Elsevie's* free academic search engine, indexes more than 30 OA repositories called "preferred Web sources" that include European institutions and grey items.

---

[23] "Library" Wikipedia, The Free Encyclopedia (accessed March 13, 2010).
[24] http://www.ekt.gr/en/index.html
[25] http://argo.ekt.gr/opac2/Help/Databases/ENU/01_Phd_EN_.html
[26] http://www.diva-portal.org/

The DART-Europe E-theses portal[27] provides access to more than 130,000 full-text research theses from 233 Universities sourced from 16 European countries (March 2010). DART-Europe, a partnership of research libraries and library consortia, is the European Working Group of the *Networked Digital Library of Theses and Dissertations* (abbrev. NDLTD, access to nearly 750,000 ETDs).

Another recent tool for resource discovery is the German *Akademisches Publikationsmanagement*[28] (abbrev. PUMA) project for the management of academic publishing.

These are but some illustrations. It is impossible to propose an exhaustive list of all European initiatives. The common point is that the notion of collection has been replaced by the concept of access. This gateway function often stays with the library; but other players enter the scene, such as publishers, search engines, computing centres etc. These new players never managed library holdings; the accent is on selection, dissemination, access, not on preservation and organization.

Roosendaal et al. (2010) depict very clearly the dynamics of this new publishing paradigm and the underlying business model. The advantage is obvious: a critical mass of information, a single access point, powerful search and selection tools, nomad access.

Some problems have been listed by Stock (2007) in her study on European ETDs in open repositories: partial or restricted access to the full text, records without full-text, missing or incomplete metadata, language barriers. Other problems are lacking standards and interoperability.

So far as we can see today, searching and collecting grey literature will not become as straightforward as it is for journals and books in the traditional publishing sector. New tools for collecting, depositing, and archiving does not make grey literature less ephemeral and volatile than in the past. Our research indicates that until an organization formulates a policy on grey literature backed by budget appropriations, the implementation of technology cannot be guaranteed and thus the environment in which grey literature has coexisted in the past will remain unstable in the likely future (Schöpfel et al., 2010).

From library to eScience

The research environment is changing and becomes more and more data-driven, with growing needs related to data acquisition, storage, processing, management, mining etc. New data integration services are already emerging, transforming data discovery on the web from lists of search results

[27] http://www.dart-europe.eu/
[28] http://puma.uni-kassel.de/

into tools that compute answers to structured questions (Fry, 2009) but as Osswald (2008) points out, so far scientific libraries have not played an important role – if at all – in e-Science projects implemented in the EU.

Access to research results (both publications and data) is "the last key ingredient of the research infrastructure (…). Thus the e-Science revolution will put libraries and repositories centre stage in the development of the next generation research infrastructure" (Hey et al., 2006).

Portals like NARCIS already include datasets in the resource offerings. With regards to the most recent developments of academic publishing (dynamic publications, 3D illustrations, primary datasets embedded in journals etc.), this is quite natural. But what is the color of datasets? Are they part of scientific literature? Or will they replace, at least partially, scientific publishing?

Today, the "article of the future" concept[29] is in the center of scientific and professional debate. Commercial publishers invest heavily in advanced editing software in order to integrate data and publication.

What about grey literature in this environment? While inclusion of raw data is a relatively new functionality for journals, supplementary material is not really new for theses and reports that often have been accompanied by CD-ROMS, maps, tables or voluminous data appendices.

Grey literature provides raw material for data mining and scientific alert services. For instance, scanning pharmacological conference announcements and abstracts allows for economic intelligence (industrial trends analysis etc.); exploitation of state of the art sections and bibliographies of PhD theses contributes to scientometrics.

The question is not if grey literature has to do with eScience but how scientific data in theses, reports, communications, working papers etc. should best be valorized. One solution is the creation of powerful data repositories by the scientific communities and their libraries, and the development of new data publishing models. Osswald (2008) warns that libraries may lose an important part of their tasks within the research community if they don't try to gain a role in eScience projects. The risk is real. But some recent initiatives provide evidence that libraries become part of the emerging scientific cyberinfrastructure. The most promising European project actually seems DataCite[30] that "promotes data sharing, increased access, and better protection of research investment".

The next step should be the interconnection between open data and publication archives, by the scientific communities and institutions[31] – if they

[29] See Elsevier's "Article of the Future" initiative http://beta.cell.com/.
[30] Project leader: TIB Hannover; http://www.datacite.org/.
[31] For instance, in a current research information system (CRIS) environment.

want to limit control of research results by commercial publishers and global information companies.

Anyway, datasets challenge the certification and preservation function of publishers and libraries. Maybe, their real place is outside of commercial distribution channels and not in the "article of the future". Tomorrow, perhaps we will not have one but many NARCIS information systems, and perhaps we will have, too, a unique gateway to access and connect Dutch, UK, German, Swedish and Czech datasets and publications. Let's dream.

The future of grey collections

This chapter tried to provide an idea of the richness and dynamics of European grey literature. Of course, it is impossible on some page to list all significant collections, such as the special collection of more than 60,000 rare publications and samizdat literature hold by the *Jagiellonian library at Cracow* or the 15,000 digital maps at the *Institut Cartogràfic de Catalunya* (abbrev. ICC) at Barcelona. The reader will find links to more resources on the websites of different LIS networks, e.g. GreyNet[32] or *Ligue des Bibliothèques Européenes de Recherche*[33] (abbrev. LIBER) with a forum on digitisation, resource discovery, heritage collections and preservation.

In the ongoing discussion on new business models of academic publishing, eScience and open access to public research results, non-commercial distribution channels will continue to play a central role as vectors of scientific communication, alongside commercial publishing.

Open archives will offer more appropriate services and functions for at least some segments of grey literature if not for all. But bibliographic control of grey literature will remain problematic despite the trend toward standardization of digital documents. And the libraries, together with their scientific communities, need to find new forms for the fundamental functions of scientific publishing, applied to open repositories, non-commercial items and datasets.

A very last remark: the article is on European grey literature, and the author is deeply attached to the European idea. But the philosophy and technology of Internet pay no attention to frontiers, nations and supranational structures. The problem is with language barriers, metadata and formats; so that the chapter closes with a plaidoyer for interoperability and standards.

[32] http://www.greynet.org/
[33] http://www.libereurope.eu/

REFERENCES:

SCHÖPFEL, J. - FARACE, D., J. *"Grey literature,"* in Encyclopedia of Library and Information Sciences, Third Edition. [online]. M. J. Bates and M. N. Maack, Eds. CRC Press, 2010, p. 2029 - 2039. Available from www: http://dx.doi.org/10.1081/E-ELIS3-120043732.

BOUKACEM-ZEGHMOURI, C. - SCHÖPFEL, J., *Document supply and open access: an international survey on grey literature*. [online]. Interlending & Document Supply, 2006, vol. 34, no. 3, p. 96-104. Available from www: http://archivesic.ccsd.cnrs.fr/sic_00181485/en/.

FARACE, D. - FRANTZEN, J. - STOCK, C. - HENROT, N. - SCHÖPFEL, J., *OpenSIGLE, Home to GreyNe's Research Community and its Grey Literature Collections: Initial Results and a Project Proposal*. [online]. In *GL10* Conference Proceedings. Tenth International Conference on Grey Literature: Designing the Grey Grid for Information Society. *Amsterdam, 8-9 December 2008*. 2008. Available from www: http://archivesic.ccsd.cnrs.fr/sic_00379643/en/.

HEY, T. - TANSLEY, S. - TOLLE, K. (Eds.) *The fourth paradigm. Data-intensive scientific discovery*. Microsoft Corporation, 2009.

HEY, T. - HEY, J., *E-science and its implications for the library community*. [online]. Library Hi Tech, 2006, vol. 24, no. 4, p. 515 - 528. Available from www: http://dx.doi.org/DOL%20%2010.1108/07378830610715383.

LUZI, D. - Di CESARE, R. - RUGGIERI, R. - CERBARA, L., *Towards an institutional repository of the italian national research council: A survey on open access experiences*. [online]. In Tenth International Conference on Grey Literature: Designing the Grey Grid for Information Society, 8-9 December 2008, 2008. Available from www: http://opensigle.inist.fr/bitstream/10068/698002/2/GL10%2^LuzLe^al%2^Conferenc^Preprin1.

MEYER, T., *Die Zentralen Fachbibliotheken und ihre Rolle fur die Fachinformation in Deutschland*. Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universitat zu Berlin, 2009, vol. 248.

OSSWALD, A., *E-science and information services: a missing link in the context of digital libraries* [online], Online Information Review, 2008, vol. 32, no. 4, p. 516 - 523. Available from www: http://dx.doi.org/10.1108/14684520810897395.

PAILLASSARD, P. - SCHÖPFEL, J. - STOCK, C., *Dissemination and preservation of French print and electronic theses*. [online]. The Grey Journal, 2007, vol. 3, no. 2, p. 77 - 93. Available from www: http://archivesic.ccsd.cnrs.fr/sic_00380488/en/.

ROOSENDAAL, H., E. – ZALEWSKA-KUREK, K. - GEURTS, P. A. - HILF, E. R. *Scientific Publishing: From vanity to strategy*. [online]. Chandos Publishing, 2010. Available from www: http://dx.doi.org/10.3152/030234207X244810.

SCHÖPFEL, J. - PROST, H., *Document supply of grey literature and open access: an Update*. [online]. Interlending & Document Supply, 2009, vol. 37, no. 4, p. 181-191. Available from www: http://dx.doi.org/10.1108/02641610911006274.

SCHÖPFEL, J. - BOUKACEM-ZEGHMOURI, C. - PROST, H., *Usage of grey literature in open archives*. In. GL11 Conference Proceedings. Eleventh International Conference on Grey Literature: The Grey Mosaic: Piecing It All Together. Washington D.C., 14-15 December 2009, December 2009.

SCHÖPFEL, J. - STOCK, C. - HENROT, N., *From SIGLE to OpenSIGLE and Beyond: An In-depth Look at Resource Migration in the European Context*. [online]. The Grey Journal, 2007, vol. 3, no. 1, p. 45-51. Available from www: http://arcm'vesic.ccsd.cnrs.fr/sic_00181592/en/.

SMITH, V., *Data publication: towards a database of everything*. [online]. BMC Research Notes, 2009, vol. 2, no. 1, pp. 113+. Available from www: http://dx.doi.org/10.1186/1756-0500-2-113.

STOCK, C. - ROCKLIN, E. - CORDIER, A. *Lara—open access to scientific and technical reports*. [online]. Publishing Research Quarterly. March 2006, vol. 22, no. 1, p. 42-51. Available from www: http://dx.doi.org/10.1007/s12109-006-0007-3.

STOCK, C., *Open access to full text and etds in europe: improving accessibility through the choice of language?* [online]. In Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape, 10-11 December 2007, 2007. Available from www: http://opensigle.inist.fr/handle/10068/697889.

TILLETT, S. - NEWBOLD, E., *Grey literature at the British library: revealing a hidden resource*. [online]. Interlending & Document Supply, 2006, vol. 34, no. 2, pp. 70-73. Available from www: http://dx.doi.org/10.1108/02641610610669769.

WILLIMSKY, J., *The Access Principle: The Case for Open Access to Research and Scholarship (Digital Libraries and Electronic Publishing)*. [online]. The MIT Press, December 2005. Available from www: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262232421.

ELECTRONIC RESOURCES:

ANRTheses [online]. Retrieved June 10, 2010 from www: http://www.anrtheses.com.fr/.

CEDA Repository [online]. Retrieved June 10, 2010 from www: http://cedadocs.badc.rl.ac.uk/.

DART-Europe E-theses Portal [online]. The DART-Europe E-theses Portal is endorsed by LIBER. c1999-2010. Retrieved June 10, 2010 from www: http://www.dart-europe.eu/basic-search.php.

DIVA. Academic Archive [online]. Uppsala University. c2000-2010. Retrieved June 10, 2010 from www: http://www.diva-portal.org/.

EThOS – Beta. Eletronic Theses Online Service [online]. The British Library Board. Retrieved June 10, 2010 from www: http://ethos.bl.uk.

GEPRIS [online]. Deutche Forschungsgemeinschaft. Retrieved June 10, 2010 from www: http://gepris.dfg.de.

INIST-CNRS [online]. Retrieved June 10, 2010 from www: http://international.inist.fr/.

NUMDAM [online]. Recherche et téléchargement d'archives de revues mathematiques numérisées. 2006 Retrieved June 10, 2010 from www: http://www.numdam.org/.

PUMA [online]. Deutche Forschungsgemeinschaft. Retrieved June 10, 2010 from www: http://puma.uni-kassel.de/.

3. The development of grey literature in the Czech Republic

Petra Pejšová

The development of grey literature up to the present day may be divided into three periods with two key turning points in the years 2005 and 2008. Before 2005, the acquisition of grey literature was oriented towards support for intra-European collaboration sponsored by the EAGLE association. Between 2005 - 2008, the European system was not in operation. Thus this period may be considered one of preparation. But it also created a gap in the centralized acquisition of grey literature. In 2008, two significant projects which have long been under preparation came into being: the *National Repository of Grey Literature*[34] (abbrev. NRGL), and the *National Register of Theses and Dissertations and a System for Seeking Evidence of Plagiarism in Theses*[35] (further referred to as "theses.cz").

The first organized acquisition of grey literature in the Czech Republic began at the start of the 1990s. This activity was initiated because of the interest of the EAGLE association in collecting grey literature from all European countries. The Czech Republic was represented in EAGLE by two bodies: the *National Technical Library* (abbrev. NTK),[36] and the *Library of the Academy of Sciences of the Czech Republic* (abbrev. AS CR Library).

As already noted in the preceding chapter, EAGLE a system called SIGLE, which catalogued the production of grey literature in Europe and provided user access. The NTK became a member of EAGLE in 1995 but had already created the *Cooperative System for Grey Literature* (abbrev. KSŠL) in 1994. This system operated on a contractual basis to assemble bibliographies of grey literature (especially dissertations) made available by cooperating Czech universities. The metadata received was converted into the specific SIGLE data format and sent in batches to *Fachinformationszentrum Karlsruhe* (abbrev. FIZ Karlsruhe) for processing. Alternatively, universities posted printed dissertations to the NTK, where bibliographies in SIGLE format were created, and sent once again in batches to FIZ Karlsruhe.

Universities were able to select the data provided based upon its relevance. As a result, the system did not catalogue the total production of dissertations in the Czech Republic. The role planned for the AS CR Library was thus never fulfilled during the existence of the SIGLE system. The NTK was the only

[34] http://nrgl.techlib.cz/
[35] http://www.theses.cz/
[36] Until July 1st, 2009, it was called the State Technical Library (STK).

active contributor and national integrator representing the Czech Republic in EAGLE.

With its 5,778 records, only a minor part of total production, the Czech Republic was still the second most active participant of the post-communist countries in the SIGLE project. The reasons for such low participation lay in the strict requirements imposed by SIGLE concerning the language of documents and records, and in the volunteer participation of Czech producers in the system. The greatest number of records, 19,524 in total, were supplied by the Russian Federation; Latvia was in the third position with 3,961 records, and the Slovak Republic was in fourth place with 850 records (OpenSIGLE, 2010).

The strength of the SIGLE system was its relatively elaborate Subject Category List (Novák, 2008). Its weakness consisted in the specific metadata format chosen, which was incompatible with MARC formats, and above all in the lack of direct links to full texts. Most member organizations solved this problem using document delivery services. However, this infrastructure did not exist in the Czech Republic at that time. Requirements for items at that time were thus satisfied using the standard *International Interlibrary Loan Service* (abbrev. IILS) with no notable connection to the KSŠL/SIGLE system.

In the final stage of its existence (2003 - 2005), EAGLE was unable to provide for the SIGLE system's capability to use the modern metadata format, or to take advantage of networking and direct electronic access to documents. The SIGLE system consequently became outdated, EAGLE's key members left and the entire association ceased to function.

The Czech KSŠL ceased activity immediately after EAGLE. It was then that the idea of building a national grey literature center was born. This center was to focus on other types of grey literature, as well, on its processing and dissemination in electronic form. In 2005, the NTK organized several meetings with representatives of universities, central libraries and other institutions focused on potential collaboration in collecting grey literature in the Czech Republic. There was evidence of interest and the NTK started work on preparing a national centre for grey literature. The NTK also actively participated in the ETD *Working Group under the aegis of the Association of Libraries of Czech Universities* (further as ETD Working Group).

These activities did not take off as quickly as had originally been anticipated. The preparatory phase ended up taking three years, lasting until 2008. During this period, bibliographies of dissertations sent by universities were persistently processed in the library catalogue of the NTK. This was also true for other commonly acquired grey literature items, such as conference proceedings, deliverables, final reports, and company literature. Grey literature was also systematically processed as part of a collaboration between AS CR institutes in the *Automated System of Publication Registration* (abbrev. ASEP), mainly conference proceedings, research

reports, dissertations, and internal literature. Other types of grey literature were not monitored by the ASEP system. All funding program outputs, especially deliverables, were stored by the respective grant agencies and ministries. Regrettably, open access for the public was only rarely provided. Grey literature was also archived on a partial basis in institutional repositories and the personal archives of outstanding personalities, and partially in catalogues of technical libraries, (dissertations, annual reports and proceedings).

The archival of ETDs at universities was a significant issue at that time. To address it, the ETD Working Group was created under the auspices of the AKVŠ in 2004, focusing on the preparation of standards and collaborative activities directed toward building a national ETD repository. Masaryk University in Brno was the most active in archiving and providing access to ETDs; it was here that the issue of plagiarism was first addressed. Other types of grey literature, however, remained neglected at the universities. These consisted of course materials and contributions to proceedings which arose in the course of extensive publication activity on the part of employees, along with technical and research reports created as part of the work of individual university departments.

Grey literature published in other research organizations, such as company or private research institutes, civil service institutions and local authorities (ministries, regions, state organizations), remained unknown. This was also true for company literature, and regularly or irregularly issued publications such as conference proceedings, catalogues or programs created as part of conferences, exhibitions and workshops. This type of grey literature entered library collections only on a partial basis, with some exceptions. For example, there were repositories of company literature in the NTK and in the Agricultural and Food Library. It is interesting to compare this to the situation abroad (Chapter 2: “Document Acquisition Strategy”).

In 2008, after the preparatory phase, systematic work on the centralized acquisition of grey literature was initiated under two projects in the Czech Republic. First was the NRGL project, headed by the NTK in collaboration with the University of Economics in Prague (abbrev. VŠE). This complex project with links to foreign partners is thoroughly documented in this chapter. The other project consists of the theses.cz system, founded by Masaryk University in Brno in collaboration with other universities. The theses.cz system assembles ETDs and searches them using a tool designed to detect plagiarism. This topic and the entire issue of grey literature at universities are described in Chapter 6 in greater detail.

The NRGL project, *The Digital Library for Grey Literature - Functional Model and Pilot Implementation*, started thanks to the support from the *Ministry of Culture of the Czech Republic* as part of research and development programs. The project is divided into three phases, lasting from 2008 to 2011. Its main goals are the systematic collection, long-term archival and

provision of access to specialized grey literature, especially to do with research and development, civil service and education, as well as from the business sphere and "open access" at the national level. To support this goal, the NTK created a functional network of partner organizations, a functional model and a pilot application. In addition, on the basis of verified technology and methods defined under the project, recommendations and standards are created for other institutions electing to build their own digital grey literature repositories. Recommendations and standards consist mainly in a recommended metadata format, exchangeable formats and templates, examples of licensing models and of legal issues resolved, preservation methodology, archival and the provision of access to digital data.

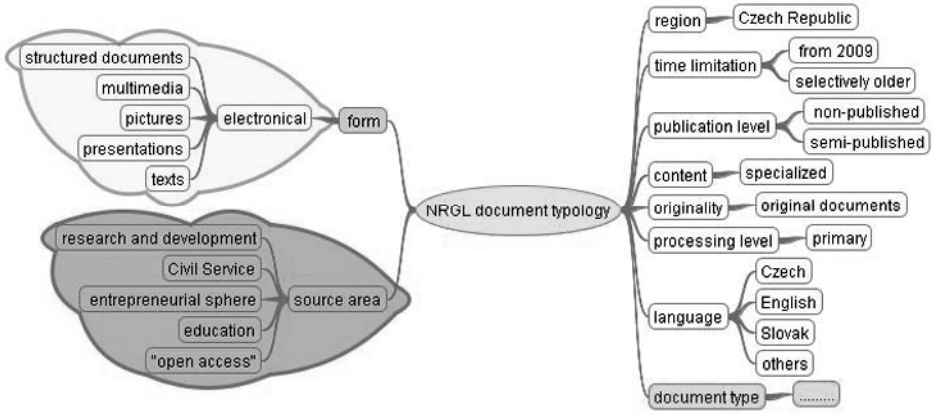

Figure 1: Basic NRGL typology (level 1)

Before collecting grey literature using the NRGL system, a typology had to be defined to clearly specify the subject of collection. Most typologies, including GreyNet, focus solely on document type. In the case of 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.

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.

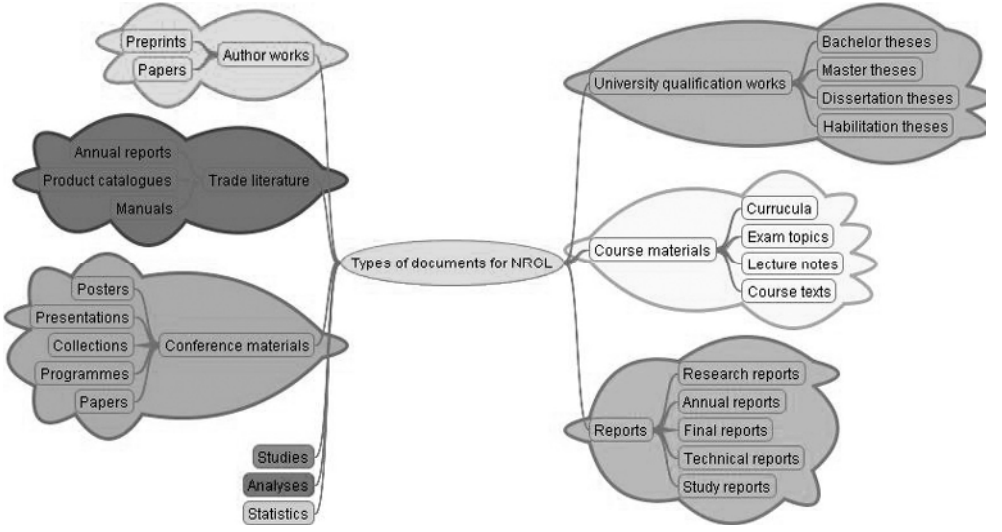

Figure 2: Typology according to the type of document (level 2)

NRGL is not intended to offer an isolated solution but rather to complement and continue other systems and projects dealing with long-term archival and the provision of access to digital documents in the Czech Republic and abroad. The significance of networking is emphasized in the introduction to Chapter 2, which contains numerous examples on building repositories and portals of grey literature.

One project which appeared before NRGL is *The Czech Digital Library* (abbrev. CDL). *The National Library of the Czech Republic* (abbrev. NL CR) created the concept of CDL – see Figure 3. The core of CDL is *The National Digital Library* (abbrev. NDL), which is depicted in the center of Figure 3. The NDL primarily focuses on documents published within the scope of protecting the cultural heritage of the Czech Republic. In contrast, NRGL concentrates on documents that are either unpublished or semi-published. In the basic functional scheme of the CDL concept, NRGL is found in the so-called standard zone, which is covered by portals and catalogues of digital documents. NRGL itself assembles metadata and full texts from local specialized and institutional repositories and other sources in the grey zone. Thus, the NTK takes over the responsibility for long-term archival and provision of access to the data acquired.

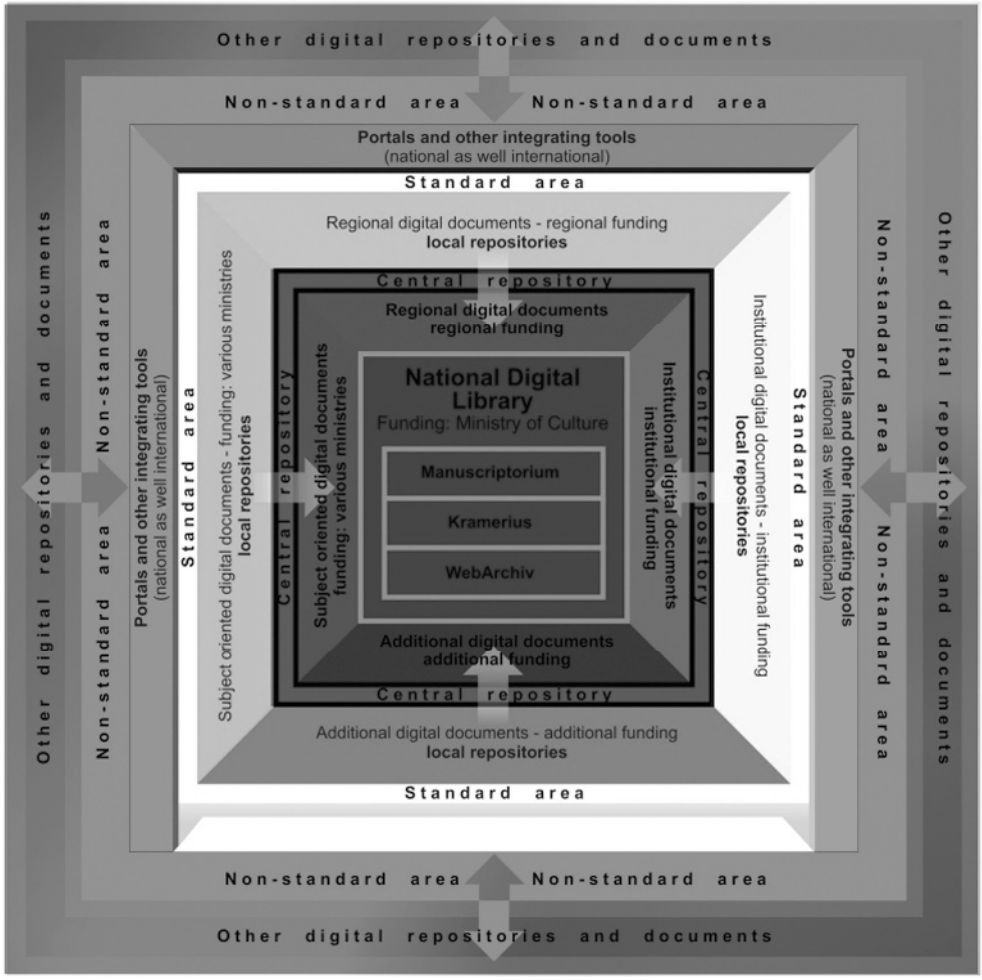

Figure 3: The Czech Digital Library (Portal of the National Digital Library, 2006) – grey literature is found in the standard zone

The institutes of the AS CR are also important grey literature producers. Grey literature is not separately monitored in the AS CR but is presented together with other scientific findings from individual departments within the scope of scientific evaluation (connected to the further funding of these research institutions). These scientific findings are monitored using the ASEP system, run by the AS CR Library. Close cooperation has been established between AS CR institutes and the library as part of networking.

Another project which should be continued under NRGL is “theses.cz”. Accumulated data could be efficiently transferred directly from the theses.cz system rather than being transferred from individual local repositories to the NRGL digital repository, designated for long-term storage. The theses.cz system could also be used as a source for the NRGL search engine without

requiring data to be harvested directly into the NRGL digital repository. Records found in NRGL would then directly refer to the documents in the theses.cz system. Such a collaborative procedure would not be able to guarantee long-term archival of data in the NRGL digital repository. Neither of these solutions is currently possible, since the theses.cz system does not use the OAI-PMH data exchange protocol and only allows the browsing of websites.

The potential for national cooperation and networking in the Czech Republic are described in Chapter 4. International collaboration is necessary in the field of grey literature to prevent the rise of individual isolated solutions at the national level. This refers to sharing standards and methodologies for processing and archival, as well as to unified access to grey literature itself.

Presently, collaboration with the GreyNet international grey literature database, based in Amsterdam, is being established. GreyNet was founded in 1992 as part of organizing for the *International Conference on Grey Literature*[37] (abbrev. GL). GreyNet aims to simplify dialogue, research and communication among researchers and organizations in the field of grey literature. After EAGLE ceased operations, GreyNet targeted for storing its functionality, especially as regards setting in place once again the cooperating network of organizations dealing with grey literature.

As already noted in Chapter 2, the French INIST revived the SIGLE system, which is currently known as OpenSIGLE.[38] A prototype of the new system was first created in 2006 based upon DSpace[39] Open Source software. At this stage, the records from FIZ Karlsruhe in Germany were converted into DSpace at the French INIST. The goal was to provide continued access to data already stored and to enable new records, including original documents, to be entered. At the end of 2008, proceedings of GLs from the last 5 years were entered (GL5 to GL9) into OpenSIGLE, including full texts and presentations. At present, the series of proceedings includes all proceedings from GL1 to GL10. Further development of OpenSIGLE is being provided by INIST, which takes care of technical requirements and produces methodologies and standards for data acquisition and links to full texts and external sources.

After creating the NRGL repository, the coordinators plan to integrate it into the projects of *Digital Repository Infrastructure Vision for European Research*[40] (abbrev. DRIVER) and *WorldWideScience.org*[41] in the second half of 2010. DRIVER is an extension for integrated European institutional open access repositories. It is a portal for searching articles, dissertations, books, reports etc., in more than 200 institutional and subject repositories in 23

[37] http://www.textrelease.com/textreleasehome.html
[38] http://opensigle.inist.fr/
[39] http://www.dspace.org/
[40] http://www.driver-community.eu/
[41] http://worldwidescience.org/

European countries and 25 languages. The WorldWideScience.org project is a global gateway to science targeting collaborative search in national and international scientific databases and portals, intended to accelerate scientific discovery and progress.

In 2008, the initial Czech websites wholly devoted to grey literature from home and abroad came into being. These websites feature continuously updated information on the NRGL project and its outcomes. These outcomes include standards (such as a metadata format, typology, data formats, data exchange format etc.) and methodologies (such as guidelines for the completion of fields in metadata format, local installation of a digital repository, maintenance of collections and modification of templates in the local installation etc.). Furthermore, these outcomes should support collaboration between NRGL and grey literature producers, for example in terms of legislation touching on grey literature. They also offer a software solution for local installation of a digital repository, downloadable from these websites. To enable broader connections, selected grey literature information sources are cited, along with references to portals supporting research and development in the Czech Republic.

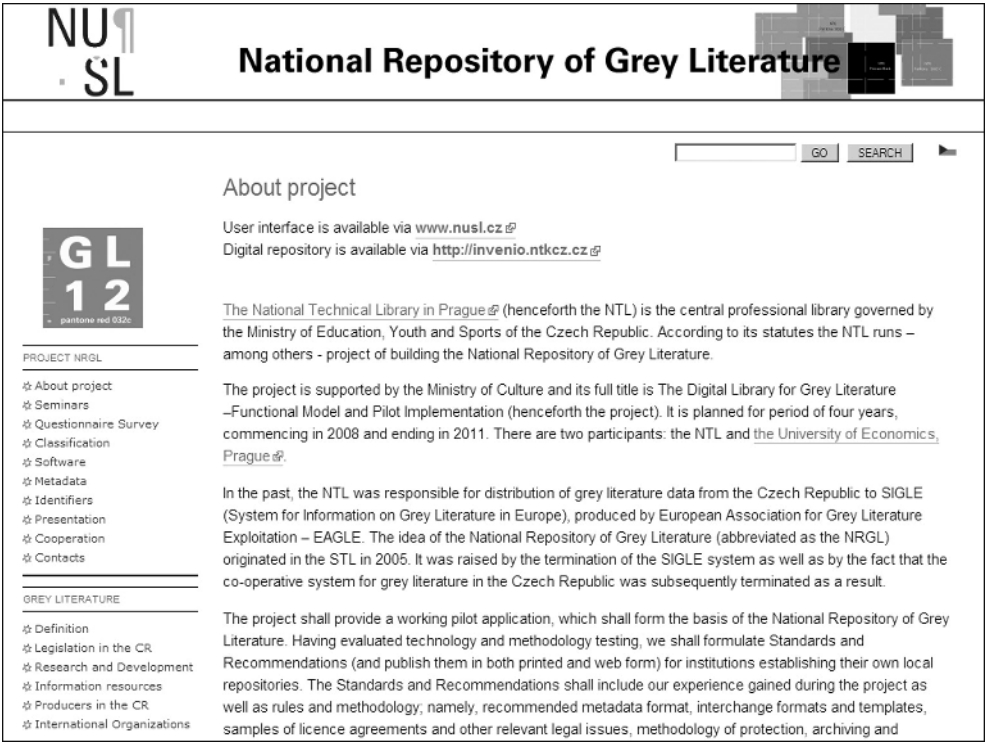

Figure 4: Screenshot of the project homepage

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 to create the site. The media wiki application supports participation in content creation within the working group and the opening of discussion groups on various topics with public participation. Since 2009, the website has also had an English version, which follows the Czech original.

Once a year since 2008, the *Seminar for the Access of Grey Literature* has been organized to inform and support scientific discussion concerning grey literature. In the workshop, updated results of the NRGL project are presented and topics relating to grey literature are discussed such as the national and worldwide status of systems dealing with grey literature, formats and standards for long-term archival, standards for describing and exchanging sources and copyright. All presentations and the full text of lectures are available in online proceedings on the project website.[42] The workshop languages are Czech and English, with proceedings held in English. The reason for this is to approach the worldwide grey literature community, which is so far scant, making each outcome in an international language precious.

Since the end of 2009, a central interface has been available to search grey literature in the Czech Republic. This central search interface offers comfortable work with the data being searched thanks to data visualization and dynamic contextual navigation. All institutions in the NRGL network are gradually integrated into this interface. The interface is available at www.nusl.cz and a detailed description of it may be found in Chapter 9.

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 of the type in question. As a result, unverified grey literature such as blogs, phone calls, discussions, and e-mails is not subject to acquisition.

Since grey literature represents a quite narrow topic, international cooperation is of immense importance. Sharing experience and providing solutions within the international community is the only way to successfully build access to grey literature. The NTK therefore collaborates closely with the international organization for grey literature, GreyNet, and with INIST, the manager of the OpenSIGLE database. At the same time, the NTK intends to provide all output from the field of grey literature in the English language, to make it understandable to the international community. In 2010, the NTK will host the GL12 conference.

[42] http://nusl.techlib.cz/sbornik/

REFERENCES:

NOVÁK, Petr. *Šedá literatura v STK: nové aktivity, nové obzory*. National Technical Library. 2008. 22 p.

ELECTRONIC SOURCES:

DRIVER [online]. Digital Repository Infrastructure Vision for Eutopean Research. Retrieved June 10, 2010 from www: http://www.driver-community.eu/.

DSpace [online]. Retrieved June 10, 2010 from www: http://www.dspace.org/.

GreyNet International [online]. Retrieved June 10, 2010 from www: http://www.greynet.org/.

INIST-CNRS [online]. Retrieved June 10, 2010 from www: http://international.inist.fr/.

OpenSIGLE [online]. INIST. Retrieved June 10, 2010 from www: http://opensigle.inist.fr/.

Portal of the National Digital Library [online]. Česká digitální knihovna - diagram - velký. October 9, 2006. Retrieved April 22, 2010 from www: http://www.ndk.cz/obrazky/czech-digital-library/.

Seminar to access of grey literature [online]. 2008. Prague: National Technical Library. Retrieved March 15, 2010 from www: http://nusl.techlib.cz/sbornik/. ISSN 1803-6015.

Textrelease [online]. Retrieved April 22, 2010 from www: http://www.textrelease.com/textreleasehome.html.

University qualification theses [online]. Faculty of informatics, Masaryk University, updated on June 10, 2010. Retrieved June 10, 2010 from www: http://www.theses.cz/.

World wide science [online]. World wide science alliance, updated on June 10, 2010. Retrieved June 10, 2010 from www: http://worldwidescience.org/.

4. The NRGL Network

Iveta Fürstová

A network of grey literature producers has come into being along with the building of the NRGL digital repository. NRGL aims to provide access to technical grey literature documents with original, valuable content. To meet this goal, operation must be established with credible partners producing quality, credible, informative grey literature. The information value of grey literature is analyzed in Chapter 1, entitled "Features and Impact of Grey Literature".

Networking partnership is open to any producer using the NRGL repository to store grey literature. Suitable partners provide grey literature containing valuable information with the potential to contribute to further development in a specific field or to heightened public awareness. NRGL's future will not be one of collecting unverified grey literature and spreading political ideas or engaging in discussions of the type seen in a blog or social network. Instead, it is oriented to specialized grey literature of a scientific, research and educational nature.

An important factor contributing to the need for networking is the necessity of abiding by Czech law, primarily laws to do with copyright, the protection of privacy and general contractual rights (Polčák, 2009). Legal opinions dictate that NRGL may freely contain metadata but a licensing contract[43] must be signed with grey literature producers to provide access to digital grey literature documents. The contract must clearly and precisely spell out what further use may be made of the documents. Grey literature is analyzed from a legal standpoint in the chapter entitled "Legal Aspects of Grey Literature".

At present, grey literature has great potential and is an important source of information. It allows research results to be distributed more quickly and more efficiently than is the case for published sources of information which rapidly lose freshness. The information potential of grey literature is enormous but it is often complicated to find the relevant material. As already noted in Chapter 1, sources other than online databases must be used and advanced methods of search are required.

When the NRGL project was in its developmental phase, there was a need to acquire information on the current state and preservation of grey literature in the Czech Republic and potential networking partners from selected areas. Before collaboration began, a questionnaire was created aim at potential

[43] A licensing contract is a contract in which the author entitles the licensee to employ the work within the scope of exercising the right (Hulva, 2007).

partners from the AS CR and public universities.[44] An online questionnaire consisting of 16 questions was available on websites and was targeted at determining the current state of grey literature, the demand for a grey literature collection and contacts for responsible persons from the institutions involved. The questionnaire served as the initial basis for collaboration with producers of grey literature.

The online questionnaire focused on just a couple of typical species of grey literature. But there is a diverse range of materials that may be classified as grey literature. In order to distinguish precisely what great literature is, several definitions are employed. These are noted in Chapter 1, "Introducing Grey Literature". Grey literature usually consists of information sources that do not pass through standard publishing procedures, but rather come from the so-called "grey zone" (Papik 2000). A NRGL typology was defined based on the questionnaire and keeping in mind the project aims and features unique to the Czech environment. Using the typology, potential partners were identified, chiefly at specialized and research institutions, schools, libraries, civil service organizations, selected companies and among specialists.

NRGL offers three possible means of collaborating with its partners. If the collaborating partner has its own database, data are transferred from this database to NRGL and then imported using a conversion table. If the networking institution does not possess its own database, the data may be stored directly in the NRGL digital repository. In this case, the partner institution will use the NRGL metadata format[45] implemented in the document import templates. The institution will have its specific collection and its own administrative rights. To administer the NRGL collection, guidelines are available in Czech, along with references to the entire documentation in English.[46] Alternatively, the collection may be administered by the NTK, which would be specified in the contract between the NTK and the organization. The third possibility is available to institutions which lack their own database but which are interested in employing the CDS Invenio system used in the NRGL project and described in Chapter 9 as a local repository. In this case, the institution will use the NRGL metadata format. The data will then be imported from the local repository directly into the NRGL repository using the OAI-PMH protocol, without the necessity for data conversion. The local installation of the preconfigured repository and detailed guidelines on its installation, setup and initial running are available at NRGL websites. All forms of collaboration are free of charge and will remain so.

[44] "The questionnaire was sent to a total of 77 respondents. 47 of them (61%) filled out the questionnaire. 18 respondents were from universities, 29 from research institutes of the ASCR."

[45] http://nusl.techlib.cz/images/Metada_NUSL_final_aktualizace2.pdf

[46] http://nusl.techlib.cz/index.php/CDS_Invenio_jako_lokální_repozitář

In support of collaboration, contracts have been created to take care of safeguarding personal data, since this data may form a part of the metadata used. A paragraph entitled Safeguarding Personal Data discusses data handling for this type of data under the contract. NTK has sought registration at the Office for Personal Data Protection to be able to process personal data in the NRGL database. After careful consideration, it was discovered that under Section 18 Subsection 1 Paragraph b) of Act No. 101/2000 Coll., the reporting obligation does not bear on the processing of personal data if it is imposed on the controller under a special Act or if such data are necessary for the exercise of rights and obligations deriving from special Acts (such as Act No. 257/2001 Coll.). It is therefore unnecessary to exercise the reporting obligation and register the NRGL at the Office for Personal Data Protection. The paragraph in question remained in the contract to protect both the provider and purchaser.

AS CR is a significant producer of quality grey literature in the area of science and research. "*Our mission and our work is science.*"[47] The mission of AS CR and its 54 institutes is to conduct basic research in a broad spectrum of natural, technical and social sciences and humanities. Information on a particular research project and why it is conducted, on research results by Czech scientists and the development and funding of science are a public matter, since AS CR is financed from the national budget. Both the academic and general public should therefore have extensive opportunities to discover and make use of these scientific results. AS CR scientists endeavor to publish most of their output in specialized national and international journals, proceedings and monographs. In the course of scientific work, however, other material may come into being which may serve as a valuable source of up-to-date information, i.e., grey literature. The NTK, in cooperation with the AS CR Library, aims to accumulate and provide access to grey literature produced by the AS CR, and thus contribute to better accessibility of this information, increasing the transparency of results of individual areas.

Long-term archival and the opportunity for easy searching of scientific grey literature brings immense advantages. This was touched on by Sophia Jones from the University of Nottingham in the second year of the *Seminar on Access to Grey Literature* (Jones, 2009). Scientists may profit from these possibilities in such a way that their work is centrally archived, its accessibility is enhanced and the results of their research are more efficiently distributed. An indubitable advantage for institutions is increasing prestige for public and funding institutions. Last but not least, long-term archiving provides public access to research results. The NRGL digital repository is a site in the electronic environment that archives and provides access to intelectual output by scientists and institutions.

[47] http://www.cas.cz/o_avcr/poslani/

During writing of the project and the design of licensing contracts, it was presumed that the AS CR institutes would have their own institutional databases. However, the majority of these institutes do not maintain grey literature databases and grey literature is not included in any other database. In practice, individual document types will be saved using the CDS Invenio or ASEP systems.

AS CR institutes produce a variable amount of diverse types of grey literature. The most common type of grey literature produced by AS CR departments are annual reports, which usually contain information on activities undertaken and institutional management for the period in question, along with research reports containing results and providing data on the current state of the research project. Annual reports are usually made public on the websites of the institutes in question.[48] Annual and research reports rank among typical types of grey literature stored in the NRGL. To save and publish such a digital document in NRGL, the institute must sign a licensing contract, even if the content has already been published on the institutional website. Only on the basis of this contract may the grey literature item be saved and published in NRGL.

The results of the questionnaire[49] indicated that AS CR institutes mainly produce reports on completed projects and university final theses.[50] Personal communications revealed that AS CR institutes run doctoral study programs under contract with universities. Dissertations written as part of doctoral studies at the university working with the AS CR institute are then defended at the faculty in question, which then stores the record as well as the dissertation itself in digital form. Therefore, only the university and not the AS CR institute has the right to enter the thesis in NRGL although the student has closely collaborated with the latter. The printed version of the dissertation is available in the library of the university, and possibly also of the institute. In addition, the AS CR is entitled to award the DSc (Doctor of Science) in the Czech Republic under a Directive of the Council for Sciences of the AS CR, No. 3/2006. The title DSc. may be awarded to candidates who have completed their dissertation thesis. These theses are archived together with dissertations; they are available in the AS CR Library and can be saved in NRGL by the AS CR. A closer survey on the production of grey literature at individual departments of the AS CR also revealed that reports on completed projects and reports on research projects are stored only in a couple of institutes. Most of these only send reports to the grant agency concerned and

[48] 90 % of the time, annual reports are available on the websites of the ASCR institutes. Research reports, however, are only available in select cases.

[49] http://nusl.techlib.cz/images/Vyhodnoceni_dotazniku.pdf

[50] 24 out of 29 ASCR institutes noted that they produce final theses and final reports on terminated projects; 15 institutes produce research reports and 6 produce technical reports.

do not archive them further. Thus, it is clear that grant agencies are another significant group of NRGL partners.

Materials from national and international conferences and workshops are also part of the grey literature documentation often found at the AS CR institutes. These include proceedings, presentations, posters, programs, lectures, information and other complementary conference materials. Proceedings occupy a position on the border between grey literature and published literature. These are non-periodical publications containing a set of individual, topic-related texts. The item may even possess an ISBN, but it is still grey literature. It is so-called semi-published literature, i.e., literature not easily accessible either in printed or electronic form. Presentations[51] and posters[52] are mostly shown in the websites of the host institution after the event; however, they are not archived. Other information and complementary conference materials are usually inaccessible and unsearchable for those not participating in the conference. Here, access to this type of documents through NRGL may be of good use.

There is a great deal of work done by authors in AS CR departments, something confirmed by a statement on the AS CR websites: "*Of over 7000 employees working in the AS CR, more than a half are scientists with university degrees.*"[53] The authors' work may be published as papers or preprints in NRGL. From an author's standpoint, a preprint is a paper which has not yet undergone the review process. From the publisher's standpoint, it is the final version of the paper which, having gone through the review process, still needs final formatting for print (Rygelova 2009). The publisher's definition of preprint is equivalent to a postprint in the database environment. For the sake of simplicity, the terms preprint and postprint are commonly classified as e-print (Harnad, 2003). Only papers or studies in preprint form can be entered in NRGL, i.e., those, for which the exclusive license[54] has not yet been granted to the publisher. As noted in the chapter entitled Legal Aspects of Grey Literature, it must be determined whether the author himself controls copyright or whether another entity exercises this function.

[51] A presentation is a public event in which the speaker shows prepared tabular supporting materials to better demonstrate the information being presented, and to heighten attention in the auditorium during the presentation. Individual pages – sheets of these materials are called slides, which can contain text, figures, tables, plots, diagrams etc.

[52] Posters are important presentation tools in conferences. They are meant to summarize scientific goals and clearly reproduce key thoughts. Posters are generally 120 cm wide and 180 cm high.

[53] http://www.cas.cz/o_avcr/poslani/

[54] Under the Copyright Act, an exclusive license means the author may not grant license to a third party and is obliged to refrain from exercising the right, unless agreed otherwise, in the same manner as that covered by the license which has been granted.

Prominent partners of NRGL in the Czech academic environment are public and private universities ranking among the most important educational and scientific institutions.[55] The general mission of universities has been aptly expressed in the status of Charles University: "*The mission of University is the propagation of education and protection of knowledge, cultivation of free thinking, of independent scientific research and artistic activity and versatile support of the genius of human society*".[56] Universities are productive suppliers of grey literature in the areas of education, science and research.

In the field of education, university final theses are the most significant type of grey literature, i.e., bachelors theses, masters theses, doctoral dissertations and habilitation theses. These are described in the chapter entitled "Access to University Final Theses". Similarly as with the AS CR, university theses should be accessible to the public, thereby contributing to their transparency and the potential to control curriculum quality. Since these species are primarily in electronic form at present (Electronic Theses and Dissertations, ETDs), information access in the digital world is quick and easy.

The legal status of final theses differs from that of other types of grey literature. Theses fall under *Act no. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act)*, which classifies this type of literature as so-called "academic work". In addition, *Act no. 111/1998 Coll., on Universities and Amending and Completing Some Related Acts (the Act on Universities)* refers to the manner of publishing theses through a digital database. All this legislation means changes both for universities and students, which must be put into practice in compliance with law. Theses are discussed from a legal standpoint in the chapter entitled "Legal Aspects of Grey Literature" in the section "Academic Work as Grey Literature". NRGL legal experts have provided an analysis of relations under the law deriving from the electronic processing and publishing of ETDs with metadata included in university databases. This should serve universities and students alike as a set of legal recommendations with reference to publication of ETDs (Polčák, 2009).

In the present moment, 14 functional university repositories are found in the registry of ETD archives on the websites of ETD Working group (Registry of Archives, 2004–2007). The most widespread system is the *Information System of Study Agenda* (abbrev. STAG) (IS STAG, c2007), developed by the University of West Bohemia in Plzeň (abbrev. UWB); since 2000 it has been used by 12 universities. For instance, Tomáš Baťa University in Zlín

[55] There are 26 public and 56 private universities in the Czech Republic.
[56] http://certik.ruk.cuni.cz/asuk/statutarni/predpisy/registrovane/statut.html

(Budínský, 2007, 2009) stores ETDs only and provides access to fundamental information on ETDs and, with author's consent, to full texts, as well.[57]

Of all available types of software systems, DSpace is often used to create digital libraries of ETDs, such as at *VŠB – Technical University Ostrava*[58] (abbrev. VŠB-TUO), *Jan Evangelista Purkyně University in Ústí nad Labem*[59] (abbrev. UJEP), *Academy of Performing Arts in Prague*[60] (abbrev. AMU), *University of Pardubice*[61] (abbrev. UPa) and, in trial, *Masaryk University in Brno*[62] (abbrev. MU). Since September 1st, 2006, MU has provided full electronic access to ETDs through the *Archive of Theses/Dissertations* in *Masaryk University Information System*.[63]

In December 2009, *Charles University in Prague* (abbrev. Charles University) opened access to the *Digital University Repository of Charles University in Prague*,[64] running under the DigiTool system of the ExLibris company. It enables long-term archival, maintainance and access to electronic documents. The DigiTool system is also implemented by *Brno University of Technology* (abbrev. BUT) to construct the *Digital Library at BUT*, with the main applications for saving and publishing ETDs as well as a digital repository (Šímová, 2009).

The ETD archive at *Mendel University of Agriculture and Forestry in Brno* (abbrev. MENDELU) is implemented as part of the integral *University Information system*, containing a public ETD registry. This archive uses a metadata file required by ETD WG. *Janáček Academy of Music and Performing Arts in Brno* (abbrev. JAMU) runs a university-wide repository accessible from the library website design for saving and accessing ETDs. Access to bibliographies and annotations is without limitation, with fulltext accessible only to JAMU students and employees.

---

[57] The STAG information system is further used by the Technical University in Liberec, Jan Evangelista Purkyně University, Academy of Performing Arts in Prague, University of Pardubice, Palacký University Olomouc, Silesian University in Opava, Ostrava University, University of Veterinary and Pharmaceutical Sciences Brno, University of South Bohemia in České Budějovice etc.

[58] http://dspace.vsb.cz/

[59] http://edu.ujep.cz/main.php?idwebu=dspace

[60] http://dspace.amu.cz/

[61] http://dspace.upce.cz/. Here, STAG has remained the interface for complete storage of metadata.

[62] http://dspace.muni.cz/

[63] The Faculty of Economics and Administration of MU runs a digital archive of final theses available at http://zp.svi.econ.muni.cz/. The archive contains bachelors and masters theses from 1994-2005, over 3500 documents total, of which 1200 are freely available to the public. Since 2006, theses have been retained within the IS MU university-wide system.

[64] http://repozitar.cuni.cz/

Since 2003, an *ETD Database* has been in operation at the *University of Ostrava* (abbrev. OU); search through saved items is possible at the *OU Portal*. Here, students save information on their theses. ETDs can be borrowed inside the reading rooms only and are found by book number from the ETD database at the OU Portal or directly from the website in question.[65] In the network of the *Technical University in Liberec* (abbrev. TUL), metadata on ETDs is accessible under the Qualification Theses link in the university-wide system.

In the case of these already established, functional university databases in which the universities themselves are archiving their ETDs, the NRGL project serves as a central searching environment. The user can find metadata and references to fulltexts of all cooperating universities in the Czech Republic in a single location. Thus, better accessibility and transparency of ETDs is achieved. Other aspects and information on ETD repositories in the Czech Republic and abroad are described in the following chapter entitled "Access to University Final Theses".

Universities are a further significant group of science and research grey literature producers, similar to the institutes of the AS CR. University employees process course materials (such as course texts and literature, lecture notes and presentations, exam topics, collections of examples, curricula etc.), reports (for instance, annual, research, project etc.), and conference materials. Pieces of authored work such as preprints and papers are also created in large numbers. The above options for collaborating with NRGL in saving, archiving and accessing this type of literature are similar as with AS CR institutes.

The grant agencies noted above also benefit by becoming part of the NRGL network. The most prominent grant agencies in the Czech Republic are those of the *Czech Science Foundation* (abbrev. GACR), and the *Grant Agency of the Academy of Sciences of the Czech Republic* (abbrev. GAAV), which prepare and implement grant projects and funding in accordance with the National Research and Development Policy of the Czech Republic. Most universities also have their own grant agencies. In addition, there is a number of ministerial grant agencies (such as the FRV of the Ministry of Education, Youth and Sports, IGA in the Ministry of Health, agencies of the Ministries of Agriculture, Ecology, the Interior and others, regional and local grant agencies (in culture and social care) and international foundations, which are a primary source. The NRGL digital repository is a suitable solution for long-term archival and access to final reports of projects funded by grant agencies.

An important link in the NRGL network is the group of civil service institutions, mainly ministries. There are presently ongoing discussions with the *Ministry of Culture, which is behind Research and Development (abbrev.*

[65] https://portal.osu.cz/wps/portal/dipl

R&D) and VISK projects. The Ministry of Education, Youth and Sports (abbrev. MEYS) should become a partner in supplying final project reports and other grey literature for NRGL. Collaboration with Czech companies (glassworks, Brewer is, industrial etc.) of interest to NRGL especially as regards annual reports, is anticipated.

NRGL could also collaborate with various personalities from fields of science and research. These partners would be able to use NRGL for long-term archival and to open their personal archives. Collaboration of this kind has already been established with Ing. Arch. Jan Moučka. During his architectural career, he has accumulated a great amount of information and a number of documents which he would like to store in NRGL, so that they will be accessible for students and scientists in the same area of research.

NRGL, which operates under the patronage of the MEYS, has as its mission the creation of a centralized environment for grey literature to unify individual institutional databases and collect grey literature items from diverse fields of science, research and education at the national level. Grey literature accessibility has recently been quite poor in the Czech Republic, especially as regards reports, conference proceedings, analyses, statistics, original work by authors etc., since items of this kind are not archived in the internal databases of scientific and research institutions. Archiving of ETDs is typically carried out at a better level of accessibility but is inconsistent; there is still no opportunity to search through the databases of all universities. The NRGL digital repository will help unify grey literature and aid in the search for faster, easier access to information.

REFERENCES:

BUDÍNSKÝ, Lukáš. Zpřístupňování e-VŠKP na Univerzitě Tomáše Bati ve Zlíně. *Konference Digitálna knižnica* [online]. 1st – 3rd October 2007, Jasná pod Chopkom, 2007. Retrieved March 15, 2010 from www: http://invenio.ntkcz.cz/record/22/files/CONFERENCE-PRESENTATION-2009-022.pdf.

BUDÍNSKÝ, Lukáš. DSpace na UTB – na půl cesty k cíli. *3. setkání českých uživatelů DSpace* [online]. 18th – 19th May 2010, Ostrava. 2010. Retrieved June 2, 2010 from www: http://dspace.vsb.cz/handle/10084/78295.

HARNAD, Stevan. Eprints. Electronic preprints and postprints. In DRAKE, Miriam A. (ed.). *Encyclopedia of library and information science* [online]. New York: Marcel Dekker, c2003. Retrieved June 2, 2010 from www: http://www.ecs.soton.ac.uk/~harnad/Temp/eprints.htm.

HULVA, Tomáš. *Právo duševního vlastnictví*. 1st Ed. Opava: Slezská univerzita v Opavě, 2007, p. 72.

JONES, Sophia. Open Access and digital repositories: the role of the DRIVER project [online]. 22nd October 2009 Prague: National Technical Library, 2009. Retrieved March 12, 2010 from www: http://nusl.techlib.cz/sbornik/2009/. ISSN 1803-6015.

PAPÍK, Richard. Zdroje šedé literatury a jejich strategický potenciál pro vědu a obchod [online]. In. *INFORUM 2000. Prague 23rd – 25th May 2000*. 2000. Retrieved March 8, 2010 from www: http://www.inforum.cz/archiv/inforum2000/prednasky/zdrojesedelit.htm.

POLČÁK, Radim. Digitální zpracování tzv. šedé literatury pro Národní úložiště šedé literatury [online]. 2009. Retrieved March 10, 2010 from www: http://nusl.techlib.cz/index.php/Právní_expertiza.

RYGELOVÁ, Pavla. Služba SHERPA/RoMEO – další život vědeckého článku. *Zpravodaj ÚVT MU* [online]. 2009. issue 20, no. 2, p. 12 – 17. Retrieved April 29, 2010 from www: http://www.ics.muni.cz/bulletin/clanky_tisk/630.pdf. ISSN 1212-0901.

ŠÍMOVÁ, Barbora. Budování Digitální knihovny VUT v Brně. *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby: 4th seminar held on 21st October 2009 at VUT in Brno* [online]. Prague: ETD Working Group within the framework of The Association of Libraries of Czech Universities. Retrieved March 15, 2010 from www: http://www.evskp.cz/seminar/2009/. ISSN 1803-7003.

Act no. 111/1998 Coll., on Universities and Amending and Completing Some Related Acts (the Act on Universities).

Act no. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), as amended.

Act no. 257/2001 Coll., on Libraries and Terms of Operating Public Library and Information Services (the Library Act).

ELECTRONIC SOURCES:

I. úplné znění statutu Univerzity Karlovy v Praze ze dne 26. dubna 2004 [online], 2004. Charles University in Prague. Retrieved March 11, 2010 from www: http://certik.ruk.cuni.cz/asuk/statutarni/predpisy/registrovane/statut.html

Qualification theses [online]. University Library. Technical University in Liberec. c2005 – 2010. Retrieved June 10, 2010 from www: http://knihovna.tul.cz/cat/Absolventske-prace-45.php.

Registry of ETD archives [online]. The ETD Working Group within the framework of The Association of Libraries of Czech Universities. c2004 – 2007, updated October 12, 2007. Retrieved March 15, 2010 from www: http://www.evskp.cz/registry.php?sek=2&tsekce=1&id=10.

ETD Database. Portal of University of Ostrava [online]. Centre for Information Technology, c2004 – 2007. Retrieved June 2, 2010 from www: https://portal.osu.cz/wps/portal/dipl.

The Digital Library of the University of Pardubice [online]. University Library of the University of Pardubice. University of Pardubice, c2002 – 2009. Retrieved June 2, 2010 from www: http://dspace.upce.cz/.

The Digital Repository of Masaryk University [online]. Masaryk University. Retrieved June 2, 2010 from www: http://dspace.muni.cz/.

The Digital Repository of VŠB - Technical University of Ostrava [online]. VŠB - Technical University of Ostrava, c2002 – 2007. Retrieved June 2, 2010 from www: http://dspace.vsb.cz/.

The Digital University Repository of Charles University in Prague [online]. Charles University in Prague. Charles University Computer Science, c2007. Retrieved June 2, 2010 from www: http://repozitar.cuni.cz/.

Doctoral Study Programmes Conducted at Institutes of the ASCR in cooperation with Institutions of Higher Education [online]. The Academy of Sciences of the Czech Republic. Centre of Administration and Operations of the ASCR, v.v.i., updated on 2nd February 2010. Retrieved March 11, 2010 from www: http://www.cas.cz/vzdelavani/doktorske_studijni_programy/prehled_akreditovanych_doktorskych_studijnich_programu.html.

DSpace - AMU repository [online]. The Academy of Performing Arts in Prague, c2008. Retrieved June 2, 2010 from www: http://dspace.amu.cz/.

DSpace - repository of qualification theses at UJEP [online]. CI Jan Evangelista Purkyně University in Ústí nad Labem, c2008. Retrieved June 2, 2010 from www: http://edu.ujep.cz/main.php?idwebu=dspace.

Fond rozvoje vysokých škol [online]. Updated April 16, 2010. Retrieved May 30, 2010 from www: http://www.frvs.cz/.

IS STAG. Information system of study agenda [online]. CIV-SIS ZČU, c2009, updated May 20, 2010. Retrieved March 15, 2010 from www: http://is-stag.zcu.cz/zajemci/.

Library [online]. Janáček Academy of Music and Performing Arts in Brno. Retrieved June 2, 2010 from www: http://www.jamu.cz/organizacni-struktura/knihovna/.

The National Repository of Grey Literature [online]. Prague: The National Technical Library, 2008. Updated on 23rd March 2009, Retrieved August 5, 2010 from www: http://nusl.techlib.cz/.

The ETD Working Group within the framework of the Association of Libraries of Czech Universities [online]. OK e-VŠKP AKVŠ ČR, c2004 – 2007, updated

October 12, 2007. Retrieved March 12, 2010 from www: http://www.evskp.cz/.

Mission of ASCR. The Academy of Sciences of the Czech Republic [online]. Centre of Administration and Operations of the ASCR, v. v. i., 2009, updated on 11th November 2009. Retrieved June 3, 2010 from www: http://www.cas.cz/o_avcr/poslani/.

Directive of the Council for Sciences of the ASCR no. 3/2006 – Rules for awarding the scientific title "Doctor of Science" at ASCR [online]. The Academy of Sciences of the Czech Republic. 2006. Retrieved March15, 2010 from www: http://www.cas.cz/miranda2/export/sitesavcr/data.avcr.cz/vzdelavani/vedecky_titul_doktor_ved/files/IN06-08.pdf.

Výzkum možností spolupráce při šíření výsledků výzkumu a vývoje. Analýza výsledků dotazníkového šetření. NTK [online]. 2008. [cit. 2010-03-10] Available from www: http://nusl.techlib.cz/images/Vyhodnoceni_dotazniku.pdf.

Basic Information on the Association [online]. The Association of Libraries of Czech Universities. c2003 – 2010, updated on 18th April 2010. Retrieved March 12, 2010 from www: http://www.akvs.cz/o-asociaci/index.html.

Qualification theses at MENDELU [on-line]. UIS MENDELU. Brno: Mendel University in Brno. [2010]. Retrieved June 2, 2010 from www: http://is.mendelu.cz/zp/.

5. Access to University Final Theses

Jan Mach

Universities stand at the pinnacle of the educational system of the Czech Republic. They are the highest centers of knowledge, independent cognition and creativity and play a vital role in scientific, cultural, social and economic development. University final theses, which demonstrate the knowledge and qualities gained by students during their studies, are defined in the Higher Education Act and represent an important component of grey literature. Universities are important NRGL partners because theses contain verified information processed under the supervision of qualified teachers and are subjected to a defense procedure. The topic of accessibility of theses in accord with the Higher Education Act has recently come under intensive discussion. Since theses are a distinctive type of grey literature, this topic is discussed separately in the following chapter.

In the Czech Republic, universities offer bachelors, masters and doctoral programs culminating in the submission of a final thesis and its defense. The bachelors degree program ends with the final state examination, usually accompanied by defense of a bachelors thesis. The masters degree program ends with the final state examination and defense of the masters thesis. In medicine and veterinary medicine and hygiene, the study ends with a state examination. The doctoral program ends with the state doctoral examination and defense of the dissertations.

The form and content of bachelors and masters is not specified in the Higher Education Act. In contrast, doctoral state examinations and dissertations are intended "to show the ability of the candidate to carry our independent activities in the area of research or development, or independent theoretical and creative artistic activities. The thesis must contain original as well as published results or results accepted for publication" (Act no. 111/1998 Coll.).

University final theses include habilitation theses submitted during the habilitation procedure, terminated with the award of the title of associate professor *("docent")*. In this case, candidates are not students submitting a thesis as part of a program of study at the university. The Higher Education Act, Section 72, defines this procedure as follows:

"(1) The aim of the habilitation procedure is to confirm the scholarly, scientific or artistic qualifications of the applicant, particularly on the basis of a habilitation thesis and its defense and of other scholarly, scientific, specialist or artistic work, as well as the applicant's competence as a teacher, based on an evaluation of his/her habilitation lecture and previous teaching experience.

(3) A habilitation thesis may be:

a) a written thesis offering new scholarly or scientific knowledge; or

b) a collection of already published scientific or scholarly papers or technical papers supplemented with a commentary; or

c) a published monograph introducing new scholarly or scientific knowledge; or

d) work of art or an artistic achievement or a collection of these, for example an outstanding public artistic activity" (Act no. 111/1998 Coll.).

University final theses, especially dissertations and habilitation theses, contain results of scientific, research, development or artistic activity and pass a defense procedure representing a process of review. This material forms a significant component of grey literature and interest in making it accessible to the public is well justified. Though these items are published to some degree, they are usually rather difficult to access. For this reason, public access to them should be better. Given the potential of modern information technology, the best way to do this is to put them in electronic form.

Until 2006, common practice was to lend these items only within reading rooms and only to students and employees of the university in question on the basis of an academic license under the Copyright Act. If the university had signed a licensing contract with students, theses were also accessible to public at large.

Act no. 216/2006 Coll., amending the Copyright Act, defined new situations for which use of the work did not interfere with copyright. University final theses fall under Section 37 of the Library license:

"(1) Libraries, archives, museums, galleries, schools, universities and other non-profit schools and educational institutions do not interfere with copyright,

(d) if originals or copies of defended masters, dissertation and habilitation theses are lent on site and this is done exclusively for purposes of research or private study, on condition such has not been prohibited by the author" (Act no. 216/2006 Coll.).

However, this still does not allow free access to theses via the internet.

A crucial legislative change took place under Act no. 552/2005 Coll., which inserted Section 47 b) into the Higher Education Act. This section imposes the obligation on universities to provide non-profit public access to dissertations, masters and bachelors which have passed defense, including comments by reviewers and the result of the defense procedure. Theses must be publicly available at least five days before the defense; they can be used for making transcripts or copies of parts or the entire document. The manner in which

theses are published in a digital database is subject to the internal regulations of each university. By submitting the thesis, the author agrees with the publishing procedure regardless of the results of the defense. The Higher Education Act does not refer to publication of habilitation theses.

Amendment of the Higher Education Act has not fully clarified the situation concerning free access over the internet. The chief problem consists in use of the term "publishing" of final theses used in the Amendment, since this is defined neither in the Higher Education Act, nor in the Copyright Act. The Copyright Act uses another term, "communicate the work to the public." For this reason, practical interpretation of this Act has differed from university to university.

Some universities still collected theses in printed form only and made them accessible in individual faculty libraries. Others made them accessible in the university library under the license defined in the Copyright Act. In this case, the reader was usually expected to become a registered member of the university library to gain access to these theses. A third group of universities made the theses freely accessible over the internet with the author's consent after thesis submission in accordance with the Higher Education Act or on the basis of a license granted by the student.

It has been several years now since the amendment came into force and university practices regarding the publication of theses continues to be varied. The number of universities offering access to theses over the internet is still growing; however, access is often subject to user registration.

Similarly as with the accessibility of theses, archiving procedures also differ significantly among universities. Each university has internal regulations governing the official discard of printed final theses. The diversity of these discard procedures is illustrated in Appendix 3 showing common practice at five selected universities.

Archival procedures have been influenced by the transition from printed final theses to electronic versions (Electronic Theses and Dissertations – ETDs). Thanks to the amended legislation, universities are not required to store printed theses within the spatial constraints of libraries and archives but may process and archive ETDs instead. For example, VŠE considers ETDs the primary document and collects and archives theses exclusively in digital form. Printed copies (if required for the defense) are returned to the author or officially discarded in the workplace in keeping with internal regulations.

In order to build grey literature repositories, long-term archival of ETDs must be assured. Such a procedure, avoiding official discarding of printed copies after a certain period, is also justified with regard to recent cases of plagiarism of final theses found several years after the candidate's graduation.

Organizations dealing with the publication of final theses

The concept of ETDs was first discussed in a meeting in Michigan in 1987 and further elaborated during the early 1990s at *Virginia Polytechnic Institute and State University* (abbrev. Virginia Tech) under Prof. Edward A. Fox, who became a co-chairman of the working group for masters theses, technical reports and dissertations. In 1996, participants in a workshop of the *Southeastern Universities Research Association*[66] (abbrev. SURA) opted for Adobe Portable Document Format (abbrev. PDF) and the Standard Generalized Markup Language (abbrev. SGML) as suitable for the presentation and archival of ETDs.

On the basis of specialized discussions and workshops, an ETD database[67] (abbrev. ETD db) originated at Virginia Tech during the same year. This software provided a comprehensive solution for the submission, processing, archival and publication of ETDs (NDLTD, c2010). The program was released free of charge internationally. Virginia Tech coordinated the development and implementation of a distributed digital library system collecting data from individual partner institutions. The system enables search by institution, creation time, author, title, keywords and the full text of theses. Fulltext of final theses was made accessible worldwide with the option to download, save and print.

In 1996, *The National Digital Library of Theses and Dissertations* group was founded in the U.S.A. When it spread to the international level, it gained its current name, *The Networked Digital Library of Theses and Dissertations* (abbrev. NDLTD). In 2003, it became a non-profit charity organization which now numbers hundreds of universities and partner institutions worldwide among its members.

The most important NDLTD document is the standard "ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations" (ETD-MS, 2008) describing final theses using 13 elements of Dublin Core and one unique element, thesis.degree, containing the sub-elements name, level, discipline, and grantor. The upgrade 1.00, revision 2, of March 21st, 2006, contains samples of metadata records in Vanilla encoding[68] and a specification for conversion into MARC-21. The ETD-MS standard served as a starting point for the proposal of other international standards, including the Czech EVSKP-MS.

Abiding by the ETD-MS standard and metadata accessibility via the OAI-PMH protocol enables ETD registries to be more efficiently built. Examples would be the search engine SCIRIUS ETD Search by Elsevier company or the

[66] http://www.sura.org/home/index.html
[67] http://scholar.lib.vt.edu/ETD-db/index.shtml
[68] http://www.evskp.cz/Seminar3/seminar3-BratkovaMach.pdf

comprehensive metadata database XTCat NDLTD Union Catalog. The latest system VTLS Visualiser,[69] offers a user-friendly interface with the option for subsequent specification of the results located.

Presently ETDs are being pursued at a number of institutions and projects at the national and international levels. An international project entitled the *DART-Europe E-theses Portal* (abbrev. DART-Europe) is especially important for the Czech Republic. This portal publishes electronic dissertations from selected European libraries (DART-Europe, c1999-2010). Metadata are automatically harvested from local repositories of partner universities. To be integrated in the system, the dissertation metadata must be accessible via the OAI-PMH protocol in Dublin Core format and complete texts must be freely available over the internet.

In the Czech Republic, librarian interest in publishing ETDs was clearly demonstrated during a presentation and subsequent discussion at the National Conference of University Libraries in Hradec Králové, held on November 19th-20th, 2003 (Prezentace, 2003). An initiative by the Club of University Librarians under the *Association of Library and Information Professionals of the Czech Republic* (abbrev. SKIP) gave rise to the ETD Working Group in 2004. The working group brought together librarians from all public universities and the NTK, which is responsible for access to grey literature at the national level.

The ETD Working Group wished to build on global experience to create the concept of a national system and unify diverse practices at individual universities, including guidelines, organizational-administrative procedures, ETD structure and ETD management. The ETD Working Group was aware of the need to improve the education of ETD authors in the field of computer science, publishing, and legislation and was ready to solve legal questions surrounding authorship. The diversity of methodologies and technologies at individual universities required preparation and implementation of standards, including the application of internationally recognized standards for filing and publishing ETDs (K otázkám, 2004).

As part of its activity, the working group prepared *A Collection of Recommendations for University Libraries*, comprising a recommended sequence of measures for introducing ETD registries at universities, collection workflow and ETD publishing, recommended university guidelines, methodologies and standards for describing related metadata. As a result, universities were ready to react to changes imposed by amendments to the Higher Education Act and the Copyright Act after 2006.

One of the crucial documents written by the ETD Working Group was entitled *Metadata Set for Electronic Theses and Dissertations in the Czech Republic* (abbrev. EVSKP-MS), created as a proposal in version 0.1 in 2005. It defined

[69] http://www.vtls.com/products/visualizer

the basic list of metadata elements for the description of ETDs defended at universities in the Czech Republic and file transmission. In addition to ETD-MS, the creators of the Czech metadata standard EVSKP-MS were inspired by the project *Digitale Dissertationen im Internet* [70] (abbrev. DissOnline), coordinated by the German National Library. Under a national project, the XMetadiss standard was prepared, containing an extensive German national metadata set, in addition to ETD-MS. This standard describes in detail dissertation and habilitation theses, including references to authors (MetaPers) and instructions for mapping the ETD-MS standard (Standard XMetadiss, 2006). On the basis of the German standard, the ETD Working Group proposed PersCZ metadata sets to describe natural persons (PersCZ, 2008) and CorpCZ to describe corporations (CorpCZ, 2008) in exchange formats, such as the metadata set EVSKP-MS and others. The proposal of EVSKP-MS was widely accepted in the academic environment, thanks to the fact that most universities were represented in the Working Group ETDs began to be collected under this standard at most universities, along with the necessary storage of descriptive metadata. The proposal was later upgraded with other administrative elements to make the current version 1.1 (EVSKP-MS, 2008).

The ETD Working Group ceased functioning in 2010, since its main goals had been accomplished. It was recommended that a new platform be developed to deal with storage strategies, long-term archival and the support of scientific and research results of universities.

ETD Repositories in the Czech Republic

As already noted, most universities opted for local ETD repositories and pertinent metadata with subsequent metadata export to library catalogues (On-line Public Access Catalog – abbrev. OPAC). The reason for this is the need for specific requirements on ETD collection and of a connection to the study agenda. The Academy of Performing Arts was among the first institutions to create its own repository using the DSpace system (OPAC Tinlib). VŠE created its own software solution for university websites, with data export to the Aleph library system.

Separation of the local repository from OPAC has the advantage of high flexibility and the option for integrating with the university information system. The main disadvantage is omission of librarians from the ETD submission process. As a result, annotations may be wrongly created (e.g., written in the first person singular or containing subjective evaluations) or keywords may be wrongly selected. The creation of subject entries by students is too advanced and therefore not recommended; EVSKP-MS metadata is primarily oriented to keywords. Subject entries based on various

[70] http://www.dissonline.de/

directed dictionaries may be entered by librarians at a later time, during metadata import to OPAC. Appendix 4 shows results of a survey on ETD publishing conducted during 2006 - 2009, performed by the ETD Working Group.

At present, most Czech universities use various solutions for ETD collection, e.g., the information study system, library catalogue, DSpace or DigiTool systems and direct ETD submission to the theses.cz system. ETDs from local repositories are mostly accessible only within the university or library. Preferred formats for collection are PDF and DOC files.

The most extensive ETD database is located at MU, which started the preparation of the concept based on the NDLTD system at the Institute of Computer Science in 2000; the project was stopped after one year. Since 2004, the MU information system has offered full support for ETD filing and has been the first to provide free public access to ETDs on their websites. The MU system uses its own ETD description.

The ETD Working Group initiated the project of building an ETD national registry. Since the Working Group could not participate in funding programs as such, the application for project funding was submitted by individual university representatives in 2006. At that time, the project was not approved due to budget limitations on centralized development projects. One year later, MU decided to solve a project similar to the ETD national registry as part of an intensive effort to combat plagiarism. Following a recommendation, seventeen universities joined in submitting the centralized development project C1/2008 from the Ministry of Education entitled the *National Registry of Theses and Plagiarism-Tracing System*. Under this project, the theses.cz system is currently in use at 29 universities. Entrance and participation in the theses.cz system will be free of charge for public Czech universities as confirmed by the coordinator and by the ETD Working Group. The theses.cz system currently contains around 100,000 ETD metadata records, with references to ETDs available either directly in the theses.cz system or in local university repositories. In addition to the EVSKP-MS standard and the OAI-PMH protocol, the system also accepts alternative means for data collection.

Metadata in the theses.cz system can be searched in full text and since 2010 has also supported the Czech declination system. Subject to the decisions of individual universities, an ETD may be accessible in one of the following regimes:

- the thesis is freely available in theses.cz,
- the thesis is available to any user after logging onto theses.cz,
- the thesis is only available to students and teachers at the pertinent university after logging onto theses.cz,

- the user is redirected to the local repository and access to fulltext is directed by the pertinent university.

At present,[71] the theses.cz system has not implemented OAI-PMH protocol for exports, which would enable it to work as a national proxy server and automatically transmit unified metadata to other national or international systems (such as NRGL, DART-Europe, and others). University representatives who want to make their data accessible in other repositories must negotiate with system managers individually and solve conversion and metadata export issues. The implementation of an open OAI-PMH protocol in theses.cz was proposed to the project coordinator by partner universities as an extension of the project.

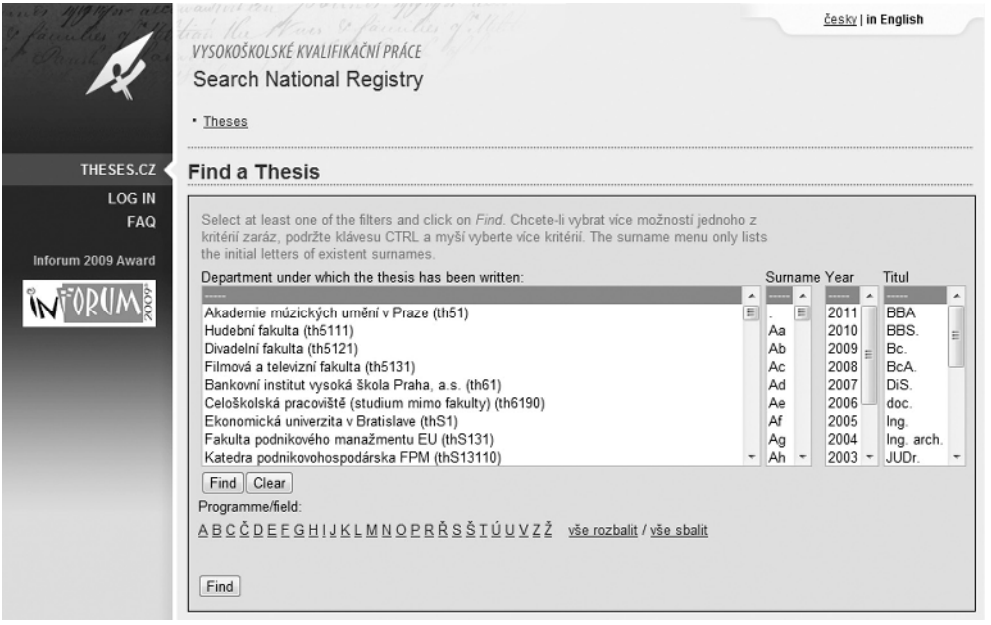

Figure 5. National Registry of Theses, www.theses.cz

The theses.cz system helps to reveal signs of plagiarism by comparing ETD texts, lectures and seminar papers. The application compares newly entered texts with existing database content and is able to detect similarities even when subtle changes are made by the author (such as changed terms or whole paragraphs of text). After a couple of hours, a PDF document is created indicating identical sections of text with reference to the source documents from the database. These reference texts are available either freely or upon request by the manager of the university in question so that signs of plagiarism may be inspected. This decision is up to the teacher or data manager of the university in question.

[71] Information updated in mid-2010.

In 2009, theses.cz system programmers built a similar system called Odevzdej.cz[72] as part of the centralized development project. This system is intended for the submission of seminar papers and should be capable of analyzing them for signs of plagiarism. Universities may use the Odevzdej.cz system directly from its website or may implement a link to their system. In 2010, the anti-plagiarism system is to be expanded to compare texts over the internet. This extension could allow the efficient discovery of student misconduct involving the copying of long extracts of text freely accessible via the internet.

Naturally these systems are unable to detect similarities if the text has been considerably altered or translated by the author. Therefore, the threat to plagiarists that their text will be freely available over the internet and thus easier to discover could be an efficient disincentive. However, only a few universities have agreed to free accessibility for their ETDs over the internet. These include Charles University, which has recently begun requiring only online registration of the user. Worries that the publishing of ETDs would result in expanded illegal copying are easily outweighed by easy access to ETDs, their integration into a central repository and the detection of similarities in texts. In contrast, the inaccessibility of ETDs among universities simplifies the situation for plagiarists and reduces the chances of uncovering the offence.

In conclusion, it should be mentioned that ETDs are an integral part of national grey literature collection under the NRGL project. The first universities to collaborate with NRGL are VŠE and Charles University. All potential partner universities can freely decide among three possible options for collaboration. The first two options offer long-term archival and accessibility for ETDs. To use these options, data must be saved in the central NRGL digital repository, either directly or via data transmission from the university repository. The third option for collaboration is integration of the university repository in the central NRGL search interface. Thus, documents may be searched but long-term archival and accessibility are not assured. All three options for collaboration are analyzed in Chapter 4. NRGL will provide searchability, accessibility, long-term data archival and, in addition, integration of universities in international registries of grey literature. This collaboration with universities has been planned since the beginnings of the NRGL project. The VSKP-MS metadata format was therefore selected as one of the chief NRGL metadata formats, as discussed in greater detail in Chapter 7.

Laws governing collecting and accessing final theses are rather specific compared to other areas of grey literature. The following chapter, entitled “Academic Work As Grey Literature”, is devoted to this topic.

[72] www.odevzdej.cz

REFERENCES:

GRULICH, Petr. *Vysokoškolské kvalifikační práce jako specifický typ archiválie a jejich digitalizace na Univerzitě Hradec Králové*. [online]. 2005. Retrieved June 10, 2010 from www: http://www.cesarch.cz/nastenka/2005/chrudim/grulich.doc.

HOROVÁ, Iva. *Aktuální stav zpřístupňování VŠKP v ČR. Výsledky průzkumu z prosince 2007* [on-line]. (PDF - 246.9kB) March 6, 2008. Retrieved June 10, 2010 from www: http://www.evskp.cz/Dokumentyver/pruzkum2007-080306104121.pdf.

HOROVÁ, Iva – KRKOŠKOVÁ, Jarmila. *Aktuální stav zpřístupňování VŠKP v ČR. Výsledky výzkumu*. [on-line] (PDF - 266.8kB) March 14, 2007. Retrieved June 10, 2010 from www: http://www.evskp.cz/Dokumentyver/vyzkum-070622150034.pdf.

K otázkám budování systémů pro zpřístupňování elektronických vysokoškolských kvalifikačních prací v České republice. Stanovisko Asociace knihoven vysokých škol [on-line]. (PDF). March 31, 2004. Retrieved June 10, 2010 from www: http://www.akvs.cz/pdf/vskpshrnuti.pdf.

VYČÍTALOVÁ, Lucie. *Aktuální stav zpřístupňování vysokoškolských kvalifikačních prací v ČR. Výsledky průzkumu z října 2009* [on-line]. (PDF). April 2010. June 10, 2010 from www: http://www.evskp.cz/Dokumentyver/pruzkum2009-100430133025.pdf.

Act no. 111/1998 Coll., on Universities and Amending and Completing Some Related Acts (the Act on Universities).

Act no. 552/2005 Coll., Amending Act no. 111/1998 Coll., on Universities and Amending and Completing Some Related Acts (the Act on Universities).

Act no. 216/2006 Coll., Amending Act no. 111/1998 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), as amended.

ELECTRONIC SOURCES:

CorpCZ: metadatový soubor pro popis korporací [online]. Zpracovatelé Eva Bratková, Jan Mach. Verze 1.0. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. May 5, 2008. 13 s. Available from www: http://www.evskp.cz/standardy/corpcz/1.0/.

DART-Europe E-theses Portal [online]. The DART-Europe E-theses Portal is endorsed by LIBER. c1999-2010. Retrieved June 10, 2010 from www: http://www.dart-europe.eu/basic-search.php.

DissOnline [online]. Digitale Dissertationen im Internet. Updated March 12, 2009. Retrieved August 5, 2010 from www: http://www.dissonline.de/.

ETD – db [online]. Digital library and archives. Retrieved August 10, 2010 from www: http://scholar.lib.vt.edu/ETD-db/index.shtml.

ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations [online]. Updated July 25, 2008. Retrieved June 10, 2010 from www: http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html.

EVSKP-MS: metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR [online]. Zpracovatelé Eva Bratková, Jan Mach. Verze 1.1. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. Retrieved July 15, 2008 from www: http://www.evskp.cz/standardy/evskp/1.1/.

Metadatový standard EVSKP-MS, v. 1.1 pro popis VŠKP a standardy související [on-line]. Eva Bratková, Jan Mach. *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby: 3. ročník semináře konaného 7. 10. 2008 na VUT v Brně* [on-line]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. Retrieved August 10, 2010 from www: http://www.evskp.cz/Seminar3/seminar3-BratkovaMach.pdf.

PersCZ: metadatový soubor pro popis fyzických osob [online]. Zpracovatelé Eva Bratková, Jan Mach. Verze 1.0. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. Retrieved May 5, 2008. 1 s. Available from www: http://www.evskp.cz/standardy/perscz/1.0/.

Prezentace. Celostátní porada vysokoškolských knihoven [online]. Hradec Králové, November 19 – 20, 2003. Retrieved June 10, 2010 from www: http://lib.uhk.cz/cpvsk2003/index.php?id=prezentace.

RIV. Rejstřík informací o výsledcích [online]. Rada pro výzkum, vývoj a inovace. Updated May 10, 2010. c2010. Retrieved August 27, 2010 from www: http://www.vyzkum.cz/FrontClanek.aspx?idsekce=956.

Standard XMetadiss. Verze 1.3. [online]. Metadata Set of the German National Library for Online. August 22, 2006. Retrieved June 10, 2010 from www: http://www.d-nb.de/eng/standards/pdf/ref_xmetadiss_v1-3.pdf.

NDLTD [online]. The National Digital Library of Theses and Dissertations. Updated March 15, 2010. Retrieved August 10, 2010 from www: http://www.ndltd.org/about/history.

VTLS Visualizer [online]. VTLS. c2008. Retrieved August 5, 2010 from www: http://www.vtls.com/products/visualizer.

6. Legal Aspects of Grey Literature

Radim Polčák

Grey Literature from a Legal Perspective

The concept of grey literature is relatively complex from a legal perspective, since existing law does not specifically treat it. The very use of the term "grey" in a legal context evokes an image of something sinister or at least of dubious nature, as in the expressions "grey imports" or the "grey economy". In actuality, however, in this case the color grey does not indicate an intermediate phase between legality and illegality but rather a specific object of legal interest exhibiting the typical features of standard literature at the same time it possesses characteristics and goals relatively novel and atypical for the law. In this case, greyness from a legal perspective should be perceived not as a sign of danger but rather as an indicator of the need for careful legal reasoning.

The standard legal model for the creation and publication of literature (which we may label "white literature") is based upon the creative activity of an author who in turn licenses a publisher to commercially present the work on the market. Such a literary work is normally disseminated by means of print, with customers paying the purchase price at the bookstore. From this, the author ordinarily derives a share of the proceeds which also, of course, finance the activities of the publisher and bookseller.

With the advent of information and communications technology, this classical model has been undergoing ever more change, not only in terms of the medium itself (commonly encountered in so-called e-books[73]) but also in terms of the economic and general transactional mechanisms for the creation and dissemination of literary works. Aside from literary works for which the author's primary intent is to "sell" the piece to a publisher or find the greatest possible market, we encounter works ever more frequently whose primary goal is to disseminate information by noncommercial means to fulfill,[74] legal or academic obligations, or even just to see the information published.

If what we have here is a relatively broad, highly metaphorical concept of grey literature, we can clarify things somewhat for subsequent legal

[73] For changes associated with the electronic publication of books, see, e.g., PICKER, R. Mediated Book, John M. Olin Law and Economics Working Paper No. 463. Chicago: The University of Chicago, 2009.

[74] On this point, cf., e.g., AUGER, Ch., P. Information Sources in Grey Literature, 2nd edition. London: Bowker-Saurn: New York, 1989.

discussion by taking the opposite tack: in place of defining grey literature in positive terms, we may substitute a description of grey literature as a genre whose meaning and goals differ from those of standard literature disseminated via a publisher. Any more precise definition of grey literature or even an attempt at its positive definition must end in defeat in the current turbulent atmosphere in this area. Even librarians themselves, whom we might consider to be the chief theoreticians of the concept of grey literature, are not completely clear on the issue.[75]

From a legal point of view it is in any event not necessary to have complete certainty about everything which might be considered to constitute grey literature. It is rather necessary to focus on the question of the kinds of legal regimes the most varied grey literature output may possess. For this it is necessary to undertake a classification of types of grey literature not according to their individual form but ideally according to the goals they achieve by their creation.[76] For now, it would appear to be adequate to consider grey literature as consisting of literary works whose primary goals include some of the following:

1. fulfillment of academic or qualification obligations,
2. reporting academic activities (mostly in research and development for the needs of the subsidy provider),
3. exchanging ideas for academic discussion (typically at conferences, in the process of writing extensive publications, during instruction etc.),
4. developing technical standards,
5. compliance with legal obligations (typically in accounting, for public tenders etc.).

In terms of entities and legal regimes for creation and dissemination of grey literature, we may then add the following typical situations to the categories introduced above:

1. a literary work created by a student or candidate (typically seminar papers, theses, habilitations, etc.) – usually academic work or work with a special regime under the higher education act,

[75] For one unsuccessful attempt to define grey literature as a term, see Gokhale, P. 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. 11/1997, Amsterdam: GreyNet, 1998, pp. 259-273.

[76] The goal is one of the central categories under applicable law. Although this may not be expressed in the text of a law, it is implicitly contained in the structure of the law and allows the law to be applied adequately in cases lawmakers did not take into consideration at the time the law was made. For more on the topic, see, e.g., HOLLÄNDER, P. Filosofie Práva. Pilsen: Aleš Čeněk, 2006, pp. 74-93.

2. a literary work created by a researcher or research team for its employer or subsidy provider, or (as typical with final reports) created by an agency which has been hired to do so – usually work for employment or work created on demand,
3. a literary work created by an independent author who self-publishes – standard individual work by an author,
4. a document created by an official or employee of a professional organization or professional association – in typical cases, there is no need to deal with protection, since normally authorial work is not involved,
5. a document in the form of a report or documentation normally created by an employee for a private or public employer – as above.

In spite of the fact that the above indicated forms of grey literature are sometimes spoken about as unpublished literature, from a legal standpoint the great majority of these works are published but often not in print by publishers. Publication of grey literature is thus carried out in particular forms depending upon the type of work (e.g. posting a final report on the web or public defence of final theses etc.) and by the increasingly frequent general form, publication in national or international databases and registers.

It is this standard form of preparing grey literature which is most problematic from a legal standpoint. The operator of the database in question does in fact publish the literary work but does not have (and does not wish to have) a legal position comparable to that of a publisher. The sense and purpose of a grey literature database is diametrically opposed to the purpose of various forms of publication by a publishing house and thereby differs greatly from the publishing house model and the legal and factual positions of the parties taking part.

General legal issues to do with the preparation of grey literature

As indicated above, out of all the legal questions to do with the creation and publication of grey literature, the focus here will be primarily on copyright. A literary work is protected in terms of copyright from the time of its creation, i.e., from the time it is put into an objectively perceivable form. The copyright on a work thus originates and is protected under law without regard to the will of the author.[77] This right includes both a moral and material component (Art.10 of the Copyright Act).

[77] The author thus need not register his work or otherwise actively advocate for its protection. The author does not even have the possibility de jure to unilaterally exclude or limit the protection of his work – see below.

The moral component of copyright is linked directly to the person of the author and includes the right to claim authorship, i.e., to be listed as the author (Art. 10 of the Copyright Act), the right to the integrity of the work (Art. 11 Par. 3 of the Copyright Act), i.e., protection against interference with the structure and content of the work, and the right to decide whether the work should be published or not (Art. 11 Par. 1 of the Copyright Act). Much more important for the acquisition and maintenance of grey literature databases is the property component of copyright, which includes the right to utilize the work in question. The catalogue of the various forms of use is relatively extensive and contains practically every conceivable type of manipulation to the work of the author, including publication in print,[78] copying etc.

The inclusion of an author's work in a closed or public database of grey literature may undoubtedly be considered use of that work for purposes of the law. For the creator or operator of the database to be able to use a particular literary work in this way, the necessary authorization must have been given. This may follow directly from the law if we are talking about free use or statutory licenses (Section 2 of the Copyright Act), or may be granted by the author or other administrator of copyright in the form of a contractual license.

In these contexts, it is important to realize that if the use of copyright works is not covered by a statutory license or by fair use, the work may only be used by negotiating an agreement with the entity holding the copyright. Such an agreement would not include a unilateral declaration or, for instance, the fact that the author has made to work freely accessible on the internet – the license must be of the character of a bilateral legal transaction. For NRGL projects this means the NTK, as the creator of the database, concludes a licensing contract with all data producers independently of the means by which the work was acquired, as is noted in the fourth chapter.

It is precisely with grey literature that the somewhat problematic situation in which the author actually cannot give unilateral agreement for further use of his work[79] has a negative impact. Particularly in the case of individual authorial works, authors have an imminent interest in disseminating their work to the maximum extent possible and would like to utilize the typical massive indirect network effects of the internet, as well. But to give an indeterminate number of interested parties the unilateral opportunity to copy and disseminate a work is impossible under the law.

[78] Here it is necessary to differentiate between the decision to publish (which is a moral right) and the publication itself (various forms are included under property rights).

[79] This is only one of the paradoxes of intellectual property rights which are being uncovered with the gradual development of the information society. See, e.g., LESSIG, L. Free Culture. New York: The Penguin Press, 2004.

A natural, albeit indirect, solution to this problem consists in so-called free licenses.[80] In technical terms, this is a public offer to conclude a license contract addressed to an unspecified range of interested parties. The fact that the work in question is offered under a free license is noted in the work itself or in its immediate proximity[81] and parties interested in copying and further disseminating the work are informed of the fact that their use is treated implicitly in the licensing terms. Depending upon the type of license negotiated, the author (or administrator of intellectual property rights) may allow further free nonprofit use of the work (this wording is normally used) or give consent for its for-profit use, as well, allow changes to the work or permit it to be incorporated in another work etc. In spite of the fact that using a free license brings with it a number of legal problems,[82] this instrument has gained significant favor, first in the USA and shortly afterwards in Europe, including the Czech Republic.[83] The significant popularity of free licenses and their notoriety in the media, combined with an ignorance of the legal context, however, often leads to situations in which a work under a free license is offered by a person not possessing the corresponding rights.

Whether the literary work in question is to be used on the basis of a free license or a standard license, it is necessary to differentiate when the intellectual property rights will be exercised by the author himself and when they will be exercised by another entity. For example, when a work has been created in an employment relationship (see below) or when the author has previously agreed an exclusive license with someone, paradoxical though it may seem, the author himself has no right to the work. An author who has published a magazine article and concluded an exclusive licensing contract with the publisher no longer has the authorization to subsequently provide his work, e.g., for processing in a database of grey literature (the magazine publisher with whom the author had earlier concluded an exclusive licensing

[80] For the fundamentals of free licenses, see, e.g., GUADAMUZ A. Open Science: Open Source Licences in Scientific Research. North Carolina Journal of Law and Technology, Vol. 7, No. 2, pp. 321-366.

[81] A standardized formulation, icons or pictograms with a reference to internet pages containing the full license agreement are used for this purpose.

[82] A free license which aims to cover various national jurisdictions in one fell swoop thus runs into differences in national legal codes, the mandatory provision of rigorous collective administration of intellectual property protection etc. For more, see, e.g., MARACKE, C. Creative Commons International The International License Porting Project. Journal of Intellectual Property. Information Technology and E-Commerce Law, No. 1, pp. 4-18.

[83] After a lengthy delay, the Czech legal system has now imported the best-known of the free licenses for literary and other media-based works, the Creative Commons license. Information on the license and a brief introduction to its use may be found at www.creativecommons.cz.

agreement would have to give permission for such use)[84] or to make the work available under a free license.

From this it follows that including an author's work in a grey literature database is only possible on the basis of a right to do so stemming directly from the law (typically with legal licenses, employee works etc.) or on the basis of a licensing contract. Unless done under one of these scenarios, it is an instance of illegal use of an author's work and the creator of the database in question may incur legal sanctions (Section 5 of the Copyright Act).

Copyright infringement is evaluated on an objective basis, using objective factual and legal criteria.[85] This means that the degree of responsibility of the infringing party (in our case, the creator of the grey literature database) is not examined, but only the fact of whether there was objective infringement of copyright.[86] It therefore does not matter whether the creator of the database acted in good faith by, e.g., mistakenly believing that the author providing an article for the database possessed the rights to it. In such a case, the creator of the grey literature database has no choice but to compensate the victim (the legitimate copyright holder) and claim damages from the one who caused the entire situation (the author or other party who illegally put the document into the database). Nor can the database creator reject or unilaterally limit responsibility to third parties.[87]

In connection with the foregoing, there is also the age-old principle of private law which states that no one can transfer more rights to another than he himself possesses. A person entering an author's work into a grey literature database who does not possess the corresponding property rights thus *de jure* cannot grant authorization for use of the work in the database to its creator. In spite of the fact that the database creator may labor under the delusion that he has been granted a valid license, in actuality he is processing and publicizing the work in question in breach of the law and is directly responsible for it to the third party suffering damages (the legitimate copyright holder).

---

[84] The reverse of this, however, is possible. In this case, the author first provides his manuscript to a grey literature database and subsequently (typically after collecting feedback and finalizing the text) decides to offer it exclusively for publication in a magazine or book. The nonexclusive license under which the original version of the article remains accessible in the grey literature database, remains valid. For more on this topic, see: Art. 47 Par. 4 of the Copyright Act.

[85] For more on the concept of strict liability, see, e.g., Havranek, J. et al. Teorie Práva. Pilsen: Aleš Čeněk, 2008, pg. 392.

[86] For more on this topic, see: TUMA, P. Copyright Act - Commentary. Prague C.H. Beck, 2007, pg. 421.

[87] Here it should be noted that the most various kinds of "disclaimers" indicating that the operator of the service is not responsible for breaches of third party rights are legally irrelevant in the great majority of cases under European law.

The only significant exception to this is a situation in which a grey literature repository is considered only as a public repository, i.e., a service which only allows users to store their data with no further manipulation taking place. If the work in question is made part of such a repository, the responsibility of the service provider is limited to cases in which he had knowledge of or could have had knowledge of infringement.[88]

Academic work as grey literature

A significant proportion of grey literature is student work. This is literary work which originates during the course of studies with the goal of fulfilling study obligations or achieving an adequate level of expertise for the granting of an academic degree. It includes the most varied kinds of ongoing study work, such as workforce seminars, course papers, as well as final theses, such as bachelors and masters theses, as described in more detail in Chapter 5.

From a regime-of-use standpoint, for student work in grey literature databases what is decisive is whether it is schoolwork simpliciter or whether it may be considered a final thesis under the higher education act.[89] School work which does not meet the legal definition of a final thesis, for instance a seminar paper, is subject to the schoolwork copyright under the Copyright Act. The student in this case possesses both moral and property rights to the work,[90] while the school may use the work, briefly put, for its own needs.[91] But the right of schools to use the work of their students does not include authorization to publish the work or disseminate it in some other way. To allow the school, for example, to include the work in its public accessible

[88] The operator of the repository in this case is not the database creator but the provider of a service for storing user data within the meaning of Art. 5 of Act No. 480/2004 Coll. For more on this type of responsibility, see POLČÁK, R. Právo na Internetu – Spam a Odpovědnost ISP. Brno: Computer Press, 2007, pg. 68 et seq.

[89] Final thesis in this context means a bachelors or masters thesis or a doctoral dissertation (see Art. 47b, Par. 1 of the Higher Education Act).

[90] In Czech academia, there is a myth that a student's teachers or even the school itself are co-authors of the work. But since student work must always be an original creation of the student himself, teachers providing supervision cannot be considered co-authors in any way. The debt he owes his teachers for creation of the work may thus be paid by thanking them but not by giving the teacher a share in the exercise of moral rights or copyright.

[91] The limits used in this case are given in Art. 35 Par. 3 of the Copyright Act. Permissible uses in such a case include using the work for instructional purposes or including it in a reference database for the monitoring of plagiarism. For more, see TELEC, I. - TŮMA, P. Autorský Zákon – Komentář. Prague C.H. Beck, 2007, pp. 380 et seq.

database of grey literature or transfer it to another repository, a voluntary licensing agreement is required.[92]

For schoolwork placed in the final thesis category under the Higher Education Act – bachelors and masters theses and doctoral dissertations – the exercise of copyright is slightly modified. There has been change in this area under the new University Act, which introduced an obligation for universities to publish final theses for availability in libraries or publicly accessible databases. For purposes of processing final theses as grey literature, most important are the provisions of the new Art. 47b Par. 1 which read as follows:

"Universities publish dissertations, masters theses, bachelor's theses and work which has been defended, including the evaluation of the examining committee and the results of the defense, in a database of qualified work which they administer. The method of publication is determined by the internal regulations of the university."

This obligation to publish final theses after defense (successful or unsuccessful) using a publicly accessible database gives implicit permission (a license) in and of itself for the university to create a publicly accessible database of final theses and process individual academic works. The scope of the institution's use of academic work is thus expanded by this provision to include a legal license to the content of final theses for the database and subsequent publication in this manner.

The implied license for final theses use carries with it, of course, the possibility for the school to provide qualification work for a grey literature database kept by third parties in the case of final theses. In such a case, however, the principle noted at the end of the last subchapter must be kept in mind, that is, that a third party cannot be given more rights than are possessed by the entity assigning them. The university is not authorized, for instance, to provide final theses for a commercial database or for commercial or nonprofit printing without permission from the author.

Brief attention should be paid in this regard to two problematic points which regularly occur in higher education, these being the publication of specifically protected information and the conclusion of licensing agreements with students. The first of these issues is especially urgent in disciplines in which final theses make use of information under special protection, such as personal data, trade secrets, know-how, classified information etc. Although the law makes no exception regarding final theses, it is clear that a mindless insistence on the letter of the law with texts containing protected information would lead to absurd consequences. Such a requirement to publish all work without exception would indirectly harm both students and

[92] In negotiating a license, the school does possess special status under a Art. 60 Par. 1 of the Copyright Act, but this, of course, does not provide authorization for direct use of the work.

the universities themselves, since it would discourage collaboration with students and universities on final theses.

A reasonable interpretation of the higher education act leads us to conclude that there are specific cases in which exceptions are justified from the requirement for publication of final theses, to the extent absolutely necessary to protect confidential information. It is thus appropriate to modify the procedure for requests to exempt work from mandatory requirements for publication or inclusion in a database of grey literature, by the internal regulations of the school in question and database technology in such a way that each request is individually justified[93] and public access to final theses is limited only to the extent absolutely necessary.

The second issue noted regarding the use of schoolwork concerns the conclusion of licensing agreements between schools and their students. Some universities require all students to sign a licensing agreement which authorizes a university to use the final thesis without restriction. In such cases, however, one may have reasonable doubts about the perfection of the contracts, especially as concerns the requirement for the contractual parties to act of their own free will. Simply put, in such cases students do not have a choice, so they sign the contract. But they do so not out of a desire to provide the school with a license to their work, but rather from fear that they would not otherwise successfully complete their studies. Contracts agreed under such circumstances may be legally defective and the final thesis is governed legally in this context by the above-noted statutory license.

Employee work as grey literature

A fundamental and significant portion of grey literature consists of highly varied types of regular or extraordinary reports on the activities of academic institutions, project teams and individual researchers. Reporting is important in the academic community not only for motivational purposes and for documenting financial transactions, but also at the professional level, where it is used to describe research results and activities.

In the great majority of cases, reports and statements are put out by employees of academic institutions or by private corporations spending research funds on research and development and then subsequently submitted by their employers to funding providers or directly to the public.

[93] There may be, for instance, a procedure calling for individual justification of requests given to the supervisor of the work and evaluated by the head of the institution or the appropriate dean. There is no reason, however, to consider removing such work from the database. Rather, public access should be limited. The presence of the work in the database, its availability upon individually justified request and its inclusion in the plagiarism control system need not be limited.

The copyright regime for such documents ordinarily falls within the scope of a special category of so-called employee works (Art. 58 of the Copyright Act).

Although with employee work, too, most moral and property rights remain with the author, the law gives the employer an exclusive legal position with regard to such work. The employer does not only deals exclusively with property rights, but also has the right to intervene in moral rights, i.e., to change or add to the work and publish it under his own name (not necessarily under the name of the author himself). University annual reports, even if put together jointly by particular academic and nonacademic employees, may be published with an author's clause designating the author simply as the university.

As is clear from the above, the author of employee work has no authorization to decide himself about the use of the work and may not offer it for publication or inclusion in a grey literature database without the knowledge of the employer. If the author does something with employee work on his own initiative, he and anyone involved with him runs the risk of legal sanction by the copyright holder, i.e., the employer.

In academia, the problem is ordinarily not with the use of employee work such as annual reports or grant reports, but rather with determining the regime for literary works arising during the course of academic activity. In this case, the focus is not on reports but on substantial literary works such as articles, textbooks and monographs created in direct connection to instructional or research activities of the author at the academic institution in question.

It would be an inappropriately broad interpretation of Art. 58 Par. 1 of the Copyright Act to state that everything an employee of an academic institution produces within his discipline constitutes employee work. The fact that the employer “acquires” copyright to everything the author creates in his discipline was also naturally be extremely demotivating for potential authors.

Although it is impossible to distinguish a clear boundary for classifying particular literary works as employee work, a reasonable solution in practice – and one which is followed by major universities – is to judge literary work as employee work only if it has been created under a directive from the employer. This category may include, for instance, syllabuses, interactive instructional materials for e-learning or texts created at work under the direction of a superior. By contrast, text books and monographs created on the basis of individual employee motivation would not be included in this category. It remains true, however, that this issue is a sensitive and complex one for the relationship between academic institutions and their employees, and requires a constructive approach by both parties which takes into account academic customs in individual research disciplines.

As the above indicates, the inclusion of employee work in a grey literature database is under the control of the employer. Even in this case, however, there are some instances in which control over this literature is limited, typically if it has come into the accompanied by obligations to an external funding provider. This may be the situation when the work in question, perhaps a research study, is the subject of a funding contract concluded between the employer (the university) and the funding provider, which has reserved to itself copyright over the resultant work. Within the context of the relationship between the university and the author, then, this is to be considered employee work. But the university is further bound to the funding provider in the manner indicated in the funding contract.[94] If, by way of example, the other insult uses the work without permission by including it in a grey literature database, he is responsible to his employer for interfering with copyright. At the same time, however, the employer is responsible for unpermitted use of the work to the funding provider which has received an exclusive license as part of the funding contract. For this reason, a partner network of institutions has been created holding copyright to employee work as part of the NRGL.

This chapter has presented a brief outline of legal issues to do with the creation and preparation of grey literature. Authorization to include literary work in a grey literature database was identified as a key issue, with the database creator often unable to verify whether the person providing the document for the database actually possesses the pertinent rights to do so. The only, albeit not ideal, defense a grey literature database operator can take against third party claims is a possible subsequent recourse.[95]

Creators of grey literature databases are often aided by the fact that copyright for grey literature is frequently not held by individuals but institutions, for school work the educational institution in question and for employee work, the author's employer. In such cases the creators of grey literature databases may contract for use of work on a group basis and the relative stability of the institutions with whom licensing agreements are concluded provides a better legal standing for any claims of recourse if work has been included in the database without authorization.

When grey literature databases allow for active access by individuals, it may be recommended from a legal standpoint that there functionality be limited to that of pure user data repositories. The operator of such a database is not in the position of its creator in the sense of Art. 88 et seq. of the Copyright Act and cannot dispose of the database or its contents in any way (for

[94] The Copyright Act labels this copyright regime "work created on demand" and gives the parties virtually a free hand in deciding how to share rights (Art. 61 Par. 1 of the Copyright Act).

[95] Grey literature databases which technically function only as user repositories are not in this situation – operators of such databases are covered by limitations on liability for information society service providers under Art. 5 of Act No. 480/2004 Coll.

example by offering it for use in a partner database), but is in a relatively safe legal position vis-à-vis potential sanctions for violating third party copyright thanks to limitations on responsibility for providers of information society services.

We have also dealt briefly with the relatively novel phenomenon of free licenses, which are closely associated with grey literature. This is a standardized offer based upon conclusion of a noncommercial licensing contract which allows authors praeter legem to provide their work to the public at large and make use of mass dissemination via the powerful direct and indirect effects of the internet. It is impossible to ignore the fact in this context that the massive appeal of free licenses and the phenomenal growth of grey literature are bringing pressure to bear for change in the basic philosophy and orientation of the current rigid – and in many cases, problematic – copyright laws.

REFERENCES:

AUGER, Ch., P. *Information sources in grey literature, 2nd edition.* London: Bowker-Saurn: New York, 1989.

GOKHALE, P. *Grey Literature Varieties - Definitional Problems, in Third International Conference on Grey Literature: Perspectives on the Design and Transfer of Scientific and Technical Information.* November 13 -14, 1997, Amsterdam: GreyNet, 1998, pp. 259-273.

GUADAMUZ, A. Open Science: *Open Source Licences in Scientific Research.* North Carolina Journal of Law and Technology, issue 7, No. 2, pp. 321-366.

HARVÁNEK, J. et al. *Teorie Práva.* Pilsen: Aleš Čeněk, 2008, pg. 392.

HOLLÄNDER, P. *Filosofie práva.* Pilsen: Aleš Čeněk, 2006, pg. 74-93.

LESSIG, L. *Free Culture.* New York: The Penguin Press, 2004.

MARACKE, C. *Creative Commons International The International License Porting Project.* Journal of Intellectual Properte. Information Technology and E-Commerce Law, No. 1, pp. 4-18.

PICKER, R. *Mediated Book.* John M. Olin Law and Economics working Paper No. 463. Chicago: The University of Chicago, 2009.

POLČÁK, R. *Právo na internetu - spam a odpovědnost ISP.* Brno: Computer Press, 2007, pp. 68 et seq.

TŮMA, P. *Copyright Act - Commentary.* Prague: C.H. Beck, 2007, p. 421.

TELEC, I. - TŮMA, P. *Copyright Act - Commentary.* Prague: C.H. Beck, 2007, pp. 380 et seq.

Act no. 111/1998 Coll., on Universities and Amending and Completing Some Related Acts (the Act on Universities).

Act no. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), as amended.

Act No. 480/2004 Coll., on Some Services of the Information Society and on Changes to Some Acts (Act on Some Services of the Information Society).

ELECTRONIC SOURCES:

Creative commons Česká republika. [online]. c2010. Retrieved June 6, 2010 from www: www.creativecommons.cz.

7. Descriptive Data for Grey Literature

Bohdana Frantíková

A format for storing metadata is an essential part of building a repository. It is therefore also requisite for building grey literature databases. Metadata is characterized by the *Czech Library and Information Science Database* (abbrev. TDKIV) as follows:

"Structured data which carries information concerning the primary data. The term metadata is primarily used in connection with electronic sources and denotes data in the broadest sense (data files, text information, image-based information, music etc.). Metadata functions for description, selection and archival. Connected to these functions, metadata is expanding for purposes of description, administration, legal claims, technical functionality, use and archival. The data is usually stored directly in the source (e.g., placed in the header information of an HTML document)" (KTD, 2003).

Although there exist several standards for metadata, metadata used for grey literature is characterized by its great variability. This variability consists not only in various formats, but also in the manner of their interpretation. This stems from the absence of instructions or rules by which particular fields are to be filled in. It may not, of course, be maintained that metadata descriptions follow no rules whatsoever but the metadata schemes to be explored (see below) lack instructions on the level of detail found in, e.g., the MARC21 manuals. In a library context, data is interpreted using cataloging guidelines (e.g., in the Czech Republic, Anglo-American cataloging guidelines). Because of this, and in spite of all the critical comments concerning the MARC21 and AACR2 formats, anyone who knows both of these standards has at their disposal metadata providing a precise description of the object being described. The problem with the metadata formats analyzed lies precisely in the absence of indication as to where the guidelines/documentation under which the data was created may be found. One way to eliminate this problem would be to have every metadata record contained a link to the appropriate documentation providing semantic help for the record. In addition to long-term availability, the rules must also allow for updating and supplementing of particular uses to take place. In June 2010 cataloging guidelines entitled *Resource Description and Access* (abbrev. RDA) were issued. The guidelines were also reviewed by a special working group of the *Dublin Core Metadata Initiative*.[96] It will be interesting to see how this rule is reflected in the processing of records for metadata.

[96] http://dublincore.org/groups/libraries/rda/

In addition to those indicated below, other schemata often encountered in digital repositories include the *Metadata Object Description Schema* (abbrev. MODS), the *Metadata Encoding and Transmission Standard* (abbrev. METS) and the *Digital Item Declaration Language* (abbrev. DIDL) (Vyčítalová, 2009). These formats, however, did not serve as a starting point for creating the NRGL format, and for this reason will not be described further.

The fundamental requirement for the NRGL format lies in achieving the greatest simplicity possible and compatibility with the Dublin Core standard. The NRGL format proposal was thus conceived to meet both these requirements.

The chief model among formats specializing in grey literature was EVSKP-MS, also based upon the Dublin Core standard and proposed for the Czech higher education environment. This format was created for eVŠKP, electronic theses, which are only one type of grey literature. It was necessary to expand the format to include further specific elements to make it suitable for qualitative descriptions and other types of grey literature. Given the significance of the EVSKP-MS format, a brief description of its components have been made part of this document. For other specific elements describing grey literature, ASEP was taken as a base, along with local bases using the *Information Register of R&D results* format (abbrev. RIV).[97]

NRGL format reflects elements of the international Dublin Core and Dublin Core Terms, ETD-MS and MARC21 formats described below. Records from the NRGL database will be put into the OpenSIGLE database, hence its characteristics will be briefly described.

NUSL base formats

Dublin Core

Dublin Core is the best known internationally recognized metadata format, primarily used for description of electronic resources. A 1995 seminar organized in Dublin, Ohio led to the beginnings of the format *Online Computer Library Center* (abbrev. OCLC).

The Dublin Core format originally arose to fill a need to improve information searches in the internet. Classic MARC library descriptions were too unwieldy in this environment, both as regards the anticipated future creation of records and with respect to their structure and semantics, which were not suited to the new environment. “Formats of the MARC are only ‘machine-readable’ but are not ‘machine understandable’, i.e., their structure and semantics are not suited to the requirements of future information

[97] http://www.vyzkum.cz/FrontClanek.aspx?idsekce=956

communication architecture, including information on information (metadata) on the internet" (Bratková, 1999).

Dublin Core is a set of metadata elements. The original format contained 13 elements but was subsequently expanded to include the current 15 elements. Later, Dublin Cora was expanded and enriched with further elements, leading to the Dublin Court Terms format. The current version of the format, "DCMI Metadata Terms", reflects changes approved and valid since December of 2006.[98]

Dublin Core does not prescribe any binding syntax (it is a semantic standard).

The Dublin Core format is extensible for particular local applications. All elements are optional and repeatable, selected data may be modified, especially for the needs of special systems via so-called qualifiers (patterns/models of data values and data types) (Bratková, 1999).

In 1998, Dublin Core metadata was recognized under internet norm IETF RFC 5013. In September 2001, the *American National Standards Institute* (abbrev. ANSI) ratified version 1.1 as national American Standard ANSI Z39.85-2007. In 2003, Dublin Core was approved by the *International Standard Organization* (abbrev. ISO) as international norm ISO 15836:2003 (Synková, 2008).

The Dublin Core format has been translated into 25 languages and is used in a number of projects and systems around the globe. Some projects have adopted it in its original form. Others, like the NRGL project, have used it as a basis to create their own format.

2010 marks the 15th anniversary of the founding of the Dublin Core Initiative, with the 10th international conference being held.[99] Conference program reflects on the past and looks to the future: conceptual models and frameworks (including RDF, DCAM, OAIS), application profiles, metadata domains (business, cultural institutions, education, government etc.), bibliographic standards (RDA, FRBR, subject entries), metadata availability and metadata for scientific data, search engines and metadata, knowledge systems and the *Simple Knowledge Organization System* (abbrev. SKOS) etc.

OpenSIGLE

16 countries took part in the development of the SIGLE system, described in the second chapter, between 1980 and 2005. In 2005, the database contained 855,260 records of research projects, final theses, conference proceedings, data files and translations in various fields of research.

---

[98] For more on these changes, see http://dublincore.org/usage/decisions/2008/dcterms-changes/ # sect-1.

[99] http://www.asis.org/Conferences/DC2010/

The INIST organization, a French member of EAGLE developed the OpenSIGLE repository on the basis of MIT's DSpace software and stored records in simplified XML format in the repository from the SIGLE database of EAGLE members who consented. Since the end of 2008, fulltext grey literature has been stored in the repository.

The OpenSIGLE repository now provides open access to more than 200 conference papers from the International Conference on Grey Literature (GL1 - GL10).[100]

More information about OpenSIGLE is given in the chapter on the Development of Grey Literature in the Czech Republic.

ETD-MS

The ETD-MS metadata format defines a set of elements for describing electronic theses. The format is not intended to replace existing local metadata formats developed for individual institutions, but rather to allow data sharing between systems. The format is based upon the elements of Dublin Core, together with specific features for eVSKP used in our format. The ETD-MS format is mentioned here because it was one of the bases for creating the Czech EVSKP-MS format on which NRGL is based.

MARC 21

The MARC21 format is managed by the *Library of Congress* in cooperation with various user groups. It is a standard for the description and representation of bibliographical information in machine-readable form and specifies data fields, field labels (tags), codes, indicators and subfields. It designates mandatory data for the exchange and sequencing of data, its corresponding codification and use guidelines.

MARC 21 was designed to facilitate the exchange of bibliographic and related information between institutions.

It is briefly described here because it is the native format for the CDS Invenio system used in NTK as a digital repository. It therefore became necessary to create a conversion table between the NRGL and Mark 21 formats. The *Conversion Table between the NRGL and Mark 21 Formats* is attached at the conclusion of this chapter.

[100] These collections are available at http://opensigle.inist.fr/handle/10068/697753.

Metadata Standard eVŠKP

The EVSKP-MS format is used to describe eVSKP. It was created and administered by the AKVŠ Commission, similarly as with the PersCz metadata format for natural persons and the CorpCz format for description of corporations. These formats were used as components in EVSKP-MS, serving for more precise entries featuring, e.g., author names and more detailed identification of the institution. It must be kept in mind when storing metadata that the data may change over the course of time. As an example, a bachelor's degree thesis may be written by someone who receives the pertinent title only after defense of the thesis.

The EVSKP-MS format is based on the Dublin Core format. On the basis of changes brought by the *Dublin Core Metadata Initiative* term definitions (abbrev. DCMI) published in January of 2008, the EVSKP-MS format was modified, as well. In addition, the format contains selected elements of the most important international standard, ETD-MS, used by the international NDLTD database. Consideration was also given in developing the format to the German national metadata standard XmetaDiss and the French metadata norm for electronic dissertations TEF 2.0 (EVSKP-MS, 2008).[101]

ASEP

The ASEP system contains bibliographical records concerning research results at institutes of the Academy of Sciences of the Czech Republic from 1985, with full records from 1993. Records are created online using a web form.

ASEP system records are also sent to the RIV database. ASEP, however, is broader in scope. Because the standards for transferring data to RIV are updated each year, the ASEP record structure is also changed. The changes are primarily to do with the addition of new fields and definitions of their contents, along with the classification of new document types.

ASEP is a format of the Marc type with fields of its own added. Data is stored in the ARL system. Information concerning the ASEP database, including a description of the format, is available at the Academy of Sciences of the Czech Republic website.[102]

[101] The format is described on the ETD Working group website at http://www.evskp.cz/standardy/evskp/.
[102] http://www.lib.cas.cz/cs/ASEP

NRGL format

The NRGL metadata format was specially created for the preparation of records of digital grey literature documents.

The NRGL metadata format uses elements of the Dublin Core format, Dublin Core Terms, EVSKP-MS, MS-EDT and custom elements.

The NRGL format is an open format which may be supplemented in the future by elements needed to operate digital libraries of grey literature and collaboration with both local data suppliers and international registries.

Changes and simplification of the original metadata concept for the NRGL format came about only after the selection of CDS Invenio as NRGL digital repository software (more information on this system may be found in the ninth chapter).

It is anticipated that any element which proves unnecessary during the pilot project will be eliminated.

NRGL format consists of mandatory and optional elements. This chapter describes mandatory elements of the NRGL metadata format in detail, specifically elements for grey literature. The entire NRGL system including optional elements may be found at
http://nusl.techlib.cz/index.php/Metadata.

Mandatory elements

In addition to attributes for individual elements, the metadata format also contains the general attribute xml:lang="[language code]" for the majority of elements. The content of this attribute is always the language code for the element in question, which has nothing to do with the independent element dc:language.

nusl:metadata

Newly created element for the NRGL format.

The root element of the metadata record for the resource. All other elements are nested within this element.

dc:identifier

This element has been converted into NRGL format from Dublin Core (dc:identifier).

Unique, permanent resource identifier.

The original aim was for this field to give a permanent identifier of the type URN:NBN, Handle etc. Currently, however, the resolver URN:NBN does not

function in the Czech Republic. It would provide and ensure a permanent connection between the identifier and the corresponding URL. The identifier for the URI record will then be generated directly within the CDS Invenio system in the format: **www.nusl.cz/ntk/nusl-ID**.

The serial number issued by the system is used for the identifier. If in the future the issue of granting permanent URN:NBN identifiers in the Czech Republic is resolved, these identifiers will be added to the records.

An identifier of the same format will also be assigned by the CDS Invenio system to the harvested records. Repeated occurrences of the dc:identifier field will be given identification numbers issued by local repositories or system numbers.

dc:title

This element has been converted into NRGL format from Dublin Core.

The main title of the resource given to it by its creator.

dc:creator

This element has been converted into NRGL format from Dublin Core.

Unnatural person primarily responsible for creating the resource content, i.e. the author of the final thesis, the chief project designer, research aim leader etc., or corporation bearing primary responsibility for the work.

dcterms:dateCreated

This element has been converted into NRGL format from Dublin Core Terms. Publication date of resource.

dc:subject

This element has been converted into NRGL format from Dublin Core.

This element contains a factual description of grey literature expressed either in the form of terms from the *Polythematic Structured Subject Heading System* (abbrev. PSH), or in the form of freely created keywords in either Czech or English. Subject classification within the NRGL digital repository is described in more detail at the end of this chapter.

dc:type

This element has been converted into NRGL format from Dublin Core.

The element is repeatable within the NRGL format. Its first occurrence indicates the typology of the NRGL document, the second occurrence indicates the type of the document under DCMI.[103]

[103] http://dublincore.org/documents/1998/10/23/type-element

First appearance: NRGL document type is the type of document according to the NRGL document typology described in the third chapter. No string should appear here (naming the type of document) which is not given in the NRGL document typology.

Second occurrence: the type of document using the DCMI typology. This element is optional under the NRGL format and may be repeated. It is ideal, in correspondence with the functions of the CDS Invenio system, if cooperating repositories respect the NRGL document typology. For older records and in taking records from other databases incapable of fulfilling this requirement, a harmonization table spanning the two typologies is created.

dc:language

This element has been converted into NRGL format from Dublin Core.

Text language of the resource.

dc:rights

This element has been converted into NRGL format from Dublin Core.

Textual information regarding rights of use. Specifies the conditions under which work may be distributed, reproduced etc.

The repeatability of this field is used in the NRGL format. The first mandatory appearance of this field always states that the full text is protected by copyright (Polčák, 2009). The second appearance of this field may indicate one of the Creative Commons licenses pertaining to the document. For use with a Creative Commons license, the Restrictions on Access field, nusl:transfer (accessRights) is automatically filled in to note the fact that the document is public. If the dc:rights field is not to be filled in, the subsequent field, Restrictions on Access, must be filled in and a document access option selected.

nusl:transfer (accessRights)

Newly created NRGL element.

So-called access restrictions to the resource are entered here which describe the manner by which the document is to be made public. The field must be filled out by worth “Public”, i.e., the file is publicly accessible via the internet without any registration requirement. This field is filled out based upon guidelines stating that other institutions may archive a document free of charge only if the document is publicly accessible. Other access options for digital documents such as domain, restricted Access, eduoram and others are not pertinent at this point.

Specific elements for grey literature

Elements for report preparation

Report preparation elements were newly created for NRGL, primarily inspired by the ASEP system. The elements were newly added to the system because existing metadata formats were not suited to specific requirements for processing documents of the report type.

nusl:resultID

Newly created NRGL element.

A repeatable field for indicating grants used to prepare the publication. The field may also indicate research aims under which the publication was prepared.

nusl:resultGrantGarant

Newly created NRGL element.

This element consists of the name of the project provider.

Elements for processing conference materials

None of the formats used by the NRGL authors as a source or starting point for creating the NRGL format contained a detailed description of conference materials. The authors therefore decided to create their own elements. In the first draft of the format, one head element was proposed with nested subelements.

Because of the system chosen for the digital repository and the fact that its internal format was Marc 21, a conversion table between Marc 21 and NRGL format was created (see the appendix entitled *Conversion Table*). In creating the conversion table, the element was simplified and an additional three elements came into being: nusl:event, nusl:eventPlace, nusl:eventDates (a more detailed description of these elements is given below). The elements as they are currently used are also suitable for working with the OpenSIGLE system, because they essentially conform to the way the system deals with conference materials. Guidelines for filling out individual fields were also brought into line with the system.

nusl:event

Newly created element for the NRGL format.

The name of the conference or other event. The official title of the conference or event is entered in the original language in full.

nusl:eventPlace

Newly created element for the NRGL format.

The country (place) in which the conference or event took place. The city and, in brackets the country in which the conference took place is entered.

nusl:eventDates

Newly created element for the NRGL format.

The date or range of dates during which the conference or other event took place, entered in keeping with the ISO 8101 norm.

Elements for preparation of EDT

The following group of components for eVSKP and course materials were converted from EVSKP-MS or the format ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. The elements concerned data on the academic title or academic degree and the granting institution, and are included in the thesis:degree element.

thesis:name

This element has been converted into NRGL format from EVSKP-MS format.

Abbreviated versions of academic titles or degrees granted by universities (e.g., PhD, doc., J.D., M.A., M. Eng. etc.).

Data on academic titles and academic degrees acquired in connection with a document being processed.

thesis:level

This element has been converted into NRGL format from EVSKP-MS format.

Type of study program under which the final thesis was prepared (doctoral, masters, bachelors etc.).

thesis:discipline

This element has been converted into NRGL format from EVSKP-MS format.

The course of study and field of study accredited by the MEYS in which the final thesis was prepared and defended.

thesis:grantor

This element has been converted into NRGL format from EVSKP-MS format.

The institution (school, institute etc.) standing behind the title granted in association with the final thesis.

Technical and administrative metadata

The CDS Invenio system archives other technical and administrative metadata in addition to bibliographical metadata. This data is not a direct component in the metadata format but serves to specify various kinds of functionality and to store administrative data. This administrative/technical data is contained in various files not essential for the actual description of the digital document. Of course, the data is important for repository administrators, which is why it is entered here.

It is followed by a list of (relevant) technical and administrative data:

PID file

The identifier is a persistent name given to the record using the syntax: nusl-ID, where ID = record ID.

SN file

The ID is an identification number assigned by the CDS Invenio system referring to the order of storage for the record in question.

doctype file

The filename, Document Type, in our case is the same as the name of the collection. The document type defines the collection in which the record is stored. The type is defined in the NRGL typology.

comboPODSBIRKA file

The name of the sub collection corresponds to the document type in the NRGL typology under which the record was stored.

FILE_INPUT file

The original name of the digital document attached to a given metadata record.

FILE_INPUT_RENAMED file

The new name assigned by the system to the digital document, replacing the original name.

indir file

The name of the directory in which the entire record, together with all attached digital documents including technical information and data, is stored.

XML metadata record

The metadata field, apart from being saved to the database (in our case, MySQL – recmysql file), is also saved as an XML file.

SuE file

Records may be stored by librarians from various institutions. The administrative metadata also includes the e-mail address of the library and saving the record.

Subject classification

In dealing with issues to do with subject classification, the authors of the NRGL format considered the following possibilities: use of the OpenSIGLE factual classification system, PSH headings and freely created keywords.

OpenSIGLE subject classification

The OpenSIGLE subject classification system corresponds yo the *SIGLE Subject Category List* (SIGLE Manual, 1999), last updated in 1999. A working analysis of the system revealed that the subject classification system in OpenSIGLE is text-only. Some categories needed updating but no expansion of the classification should be expected in the immediate future. The OpenSIGLE classification system is divided into 20 basic categories which contain subcategories. The number of these various for individual categories, with the smallest number being 3 (Category 120 Mathematical Sciences, General) and the greatest number being 26 (four Category 060 Biological and Medical Sciences, General).

PSH

The Polythematic Structured Subject Heading System is Czech-English dictionary divided into 44 thematic groups with hierarchically organized headings. It is designed for subject indexation of documents and subsequent search. PSH is normally published in SKOS format, which allows access to the entire headings system in machine-readable form.

The PSH dictionary is used as part of NRGL for the following reasons: it was developed within the NTK framework, is bilingual and suited for the subject description of electronic information resources.

The PSH dictionary is used in the NRGL format down to second-level entries. Implementation of the entire dictionary in the CDS Invenio system is under preparation.

Automatic indexation of fulltext headings from the PSH using the Maui Indexer is planned. “Assigning headings from the controlled vocabulary functions as follows: First, a model is created requiring fulltext terms at the entry point from the dictionary being used. The Maui Indexer will *learn* the proper indexing from this. If this model is available, new documents may

begin to be processed which have not been indexed. The Maui Indexer analyzes the content and relationships in the documents and the probabilities captured in the model are used to compute the best possible terms for indexing from the dictionary being used" (Mynarz, 2009).

Mapping the PSH to the OpenSIGLE classification is done outside the repository in exporting to the database.

Freely created keywords

Because the PSH dictionary is implemented only to the second heading level and the records contain various types of descriptions, there is also a variant allowing the use of freely created keywords or subject terms from other systems.

REFERENCES:

BRATKOVÁ, Eva. Metadata jako nový nástroj pro komunikaci webovských informačních zdrojů. *Národní knihovna Knihovnická revue* [online]. 1999, issue 10, No. 4. Retrieved April 19, 2010 from www: http://full.nkp.cz/nkkr/pdf/9904/9904178.pdf. ISSN 1214-0678.

KRČÁL, Martin. O Komisi: základní informace [online]. Asociace knihoven vysokých škol ČR, Odborná komise e-VŠKP. Updated April 14, 2008. Retrieved May 27, 2010 from www: http://www.evskp.cz/komise.php?tsekce=6&sek=.

MYNARZ, Jindřich. Jak lze prakticky využít Polytematický strukturovaný heslář pro věcný popis elektronických zdrojů. *Ikaros* [online]. 2009, issue 13, No. 12. Retrieved June 6, 2010 from www: http://www.ikaros.cz/node/5872. URN-NBN:cz-ik5872. ISSN 1212-5075.

POLČÁK, Radim. Digitální zpracování tzv. šedé literatury pro Národní úložiště šedé literatury [online]. 2009. Retrieved March 3, 2010 from www: http://nusl.techlib.cz/index.php/Právní_expertiza.

SCHÖPFEL, Joachim - STOCK, Christiane - HENROT, Nathalie. From SIGLE to OpenSIGLE: An in-depth look on resource migration in the European context. In *Eighth International Conference on Grey Literature: Harnessing the Power of Grey, New Orleans, United States, 4-5 December 2006*. (Conference Program and Abstracts: GL-conference series, ISSN 1385-2308; No. 8). Preprint conferencial material. Available z www: http://opensigle.inist.fr/bitstream/10068/697774/2/GL8,%20Schopfel%20et%20al.,%202007,%20Conference%20Preprint.pdf.

SIGLE Manual. Part 2. System for Information on Grey Literature in Europe. European Association for Grey Literature Exploitation. EAGLE Technical Committee: April 1999.

SYNKOVÁ, Veronika. Vybrané aplikace metadatového formátu Dublin Core. *Ikaros* [online]. 2008, issue 12, No. 5. Retrieved April 19, 2010 from www: http://www.ikaros.cz/node/4715. URN-NBN:cz-ik4715. ISSN 1212-5075.

VYČÍTALOVÁ, Hana. Foreign digital archives of research and technical reports and perspectives of their integration. Prague 2009. 112 p. Master thesis. Charles University in Prague, Faculty of Arts and Philosophy, Institute of Information Studies and Librarianship 2009. Supervizor of the master thesis is Eva Bratková.

Act No. 130/2002 Coll., on the Support of Research and Development from Public Funding and on Changes to Some Associated Acts (the Act on Support of Research and Development).

ELECTRONIC SOURCES:

Dublin Core Czech [online]. Masaryk University in Brno. Updated November 18, 2010. Retrieved June 10, 2010 from www: http://www.ics.muni.cz/dublin_core/index.html.

Dublin Core Metadata Initiative. 2008. *DCMI Metadata Terms* [online]. Dublin Core Metadata Initiative. Retrieved May 27, 2010 from www: http://dublincore.org/documents/dcmi-terms/.

EVSKP-MS: metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR [online]. Compilator: Eva Bratková, Jan Mach. 1.1. Version. Prague: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR, July 15, 2008. Retrieved May 9, 2010 from www: http://www.evskp.cz/standardy/evskp/1.0/.

KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Prague: National Library of the Czech Republic, 2003. Available from www: http://sigma.nkp.cz/cze/ktd.

Academy of Sciences Library. 2007. Evidence publikační činnosti v AV ČR: Manuál [online]. Prague: Academy of Sciences Library, [c2007]. Retrieved May 23, 2010 from www: http://www.iach.cz/knav/manual/manual.htm.

Metadatový formát NUŠL. Compilator: Veronika Synková, Radka Římanová a Bohdana Frantíková. Beta 0.3. Version. Prague: National Technical Library, 2009.

RIV. Rejstřík informací o výsledcích [online]. The Research and Development Council. c2010. Updated April 10, 2010. Retrieved August 27, 2010 from www: http://www.vyzkum.cz/FrontClanek.aspx?idsekce=956.

Appendix **Conversion Table between NRGL and MARC 21 Formats**

This document includes a conversion table for converting between elements in NRGL and MARC 21 formats. The need to prepare a conversion table as part of the project arose in connection with the choice of software for NRGL, the CDS Invenio system, whose native format is MARC 21.

The starting point for preparing the conversion table was the *Library of Congress* (Network Development and MARC Standards Office) Dublin Core conversion to MARC 21, available at http://www.loc.gov/marc/dccross.html. Elements in NRGL format taken from the ETD-MS metadata standard were converted in the manner indicated on the NDLTD website at http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html #thesis.degree where allowed by the CDS Invenio system.

For non-Dublin Core elements, we have attempted to find a corresponding field in MARC whenever possible.

NRGL element	MARC 21 - Identifier, field, subfield	Note
nusl:metadata (NRGL Metadata Record)		not assigned
dc:identifier (source identifier)	856 40$u	
dc:title	245 00$a	
dc:title (nusl:typeTranslated)	245 00$b	
dcterms:alternative	246 33$a	
dcterms:alternative (nusl:typeTranslated)	246 33$b	
dcterms:alternative (nusl:titleType=volTitle)	246 30$p	
dcterms:alternative (nusl:titleType=volNumber)	246 3#$n	
dc:creator	720 1#$a (person) 720 2#$a (corporation) 720 ##$a (unspecified)	We convert to field 720 because we have created neither main nor secondary headings for corporations or persons.
dc:contributor	720 1#$i (person) 720 2#$i (corporation) 720 ##$i (unspecified)	We convert to field 720 because we have created neither main nor secondary headings for corporations or persons.
dc:contributor.role	720 $e	
nusl:event	711 2#$a	
nusl:eventPlace	711 2#$c	
nusl:eventDates	711 2#$d	
thesis.degree.name	502 $b	

thesis.degree.level	502 $g	
thesis.degree.discipline	656 #7$a	
	656 $2	(by AKVO keys)
thesis.degree.grantor	502 $c	
nusl:resultID	C12$a project number; C13$a research aim (aim code)	
nusl:resultGrantGarant	C12$b CEP provider; C12$c other provider	
dc:publisher	260 ##$b	
dcterms:dateCreated	046 ##$b	Special Coded Dates/Date Created
dcterms:dateSubmitted	502 ##$a	(Dissertation Note) with initial label “Date Submitted”
dcterms:dateAccepted	502 ##$a	(Dissertation Note) with initial label “DateAccepted”
dcterms:modified	046 ##$j	Special Coded Dates/Date Modified
dc:subject	650 #7$a	
	650 $2	PSH
dcterms:abstract	520 ##$a (Summary, etc. note)	
dc:type (nusl:typeType)	655 #7$a	
	655 $2	NRGL document type
dc:type (nusl:typeType)	655 #7$a	
	655 $2	DCMI type
dcterms:medium	340 ##$a (Physical Medium)	
dc:language	041 07$a	
dcterms:BibliograficCitation	500 ##$a	
dc:rights	540 ##$a (Terms Governing Use and Reproduction Note)	
dcterms:extent	300 ##$a (Physical Description)	
dcterms:available	307 ##$a (Hours, Etc.)	
nusl:dateDelivered	008__00-05	
nusl:modified	005	
nusl:transfer (accessRights)	506##$a (Restrictions on Access Note)	

8. Analysis and supporting documents for the selection of software for NRGL

Petr Karlach

One of the most important steps in implementing the NRGL project was the selection of software for content storage, administration and access of the grey literature database. But the selection criteria included more than just the grey literature repository itself. Software was chosen to enable use of cooperating digital repositories under NRGL, as well.

The choice of software for the digital library was essentially no different than choosing software for any other activity. Although it may seem that software would be the cornerstone of a planned digital repository because of its nature, in the end the software is only a tool which enables the repository to be implemented. The most important preparatory phase for the selection of software, then, is a good description of the repository itself, along with its goals, chief functions, processes, activities, content and ties to its environment. With a well-defined repository, it becomes easy to deduce the chief requirements for the software system and the criteria for its choice. In contrast to the commercial environment, digital repositories are largely operated by nonprofits and Open Source software (software with available source code, normally obtainable without licensing fees) is characteristic, with several of the leading examples having already been used in installations from which it is possible to learn.

This chapter is divided into two sections. The first section focuses on defining functional requirements for the grey literature repository used as a basis for the NRGL software tender. The second section describes and analyzes Open Source systems selected using designated criteria. The analysis is available on the NRGL project website to serve as inspiration for potential suppliers taking part in the public tender for NRGL software. The outcome of the tender is described in the introduction to the following chapter.

With a quality definition of the repository itself in hand, it becomes possible to extrapolate selection criteria. As regards background material for NRGL software selection, criteria were designated iteratively in several steps and during the initial proposal phase were divided into five groups for ease of comprehension. Without insisting on a precisely relevant grouping, the associated requirements were as follows:

Digital library content management – this criterion covers the chief functions of the digital library, including the creation, storage, control and versioning of content (e.g., when migrating files), digital libraries, linear (list) and hierarchical searching, metadata management, various data acquisition mechanisms (including harvesting - OAI-PHM, Web Services)

interoperability for collaboration with other digital libraries (e.g., Z39.50, OAI-PMH, SRU), support for large volumes of data, multilingual search of metadata and files (fulltext), logical and proximal search, support for different file formats – text (e.g., DOC, RTF, PDF), image files (e.g., TIFF, JPEG, GIF, PNG), presentations (e.g., MS PPT, PDF and PostScript), structured data (HTML and XML), audio and video (e.g., WAV, MP3, AVI, MPEG4, RealAudio and video), encoding support (e.g., ASCII, Unicode, UTF-8).

User Interface - internet access with support for the major browsers, the potential to adapt the user interface for different user roles inside and outside the library, a multilingual user interface, multilingual presentation of search results, web 2.0 features.

Administration and user access control - classification and profiles of users based upon their roles (e.g. RBAC), user authentication based upon usernames and passwords, controlled IP address access, access via proxies, monitoring and reporting to evaluate user behavior for later use in improving the library's digital services, user authentication (e.g. Shibolet, Edurom, OpenID), recording user activities for billing purposes.

System Administration - this requirement focuses on the digital library operating environment, which can be difficult and demanding, especially for very large digital libraries. Included are tools for managing metadata, setting automatic data collection mechanisms, indexing, automatic keyword generation, output formatting etc.

Other requirements - user and administrator documentation, helpdesk (support from the system manufacturer/installer), extra functionality for users such as discussion and communication groups, automatic notification of various events (the appearance of new documents meeting certain criteria), RSS output, user mailboxes and comments on documents, requirements for operating hardware and software etc.

Upon concluding this step and evaluating the initial collection of criteria, we returned once again to the NRGL functional specification. With knowledge of the basic software requirements, we formulated a brief description of the main process and main activities of NRGL for purposes of proceeding further:

Content generation - content providers make content available (both metadata and the files themselves) in various ways, via web services, standard OAI-PMH, e-mail, FTP data transfer, on media and by direct transfer. If metadata acquired through OAI-PMH includes the URL for the file(s) and there is an agreement with the provider so permitting, the files themselves may be downloaded.

Content processing - the data thus provided must be edited, metadata must be converted into NRGL format (the software must enable definition of the actual metadata structure) using conversion templates, and the files

converted into presentation format (PDF/A), batches prepared for import into the database with the possibility of controls, metadata and presentation data must be stored in the database – potentially including the original file in various formats – indexes and keywords must be generated and search output must be formatted.

Access rights - users access content via a web interface anonymously as public users or, if access is being made from particular institutional addresses, the system will sign them in as anonymous institutional users and automatically assign the appropriate role, or they may sign in using a name and password issued by the library system, or may gain authorization using systems such as Shibolet, Edurom, or OpenID. Users may view content with access rights dependent upon their roles based upon various criteria linearly using lists, as well as hierarchically using a tree structure, or in Web 2.0 form by searching metadata as well as content using Boolean expressions, fulltext or proximity. Searching and sorting must take into account the nature of the Czech language, including not relying on diacritics. They may save search results in permanent user storage boxes (shopping carts), write commentary on the content by means of discussion groups, request notification by e-mail or to their desktop on events in the collection (new records etc.). The system enables user activity to be recorded for purposes of billing or accessing paid services using an interface to another system.

Maintenance and protection of content – quality controls on content in terms of content management, redundant data, metadata quality including the integrity of restrictions, formulation of queries on the database, checks on primary files for sustainability of content etc.

System administration - administrative tools for managing, configuring and administering the entire system, including managing operations on the central database and their timing with the option for manual intervention (e.g., shifting time-intensive tasks underway to a later period), managing user accounts, roles and access rights, automatic notification of various events, the possibility to formulate administrative queries on the database, a data backup system.

A general functional specification of the digital repository software was constructed on the basis of this brief description of NRGL and the initial software requirements. The following functionality was required:

- a system and database for storing a large number of library records based upon the actual metadata definition from various internal and external sources and their rapid, effective retrieval provided to end users via an internet user interface which supports interoperability with other digital document repositories
- automatic testing of the quality and integrity of metadata (e.g., field length, numeric/alphanumeric content, mandatory/optional/dependent

fields) and reporting on any inadequacies to administrators or cataloguers

- automatic extraction of keywords from text documents based on the frequency of their appearance in the dictionary
- metadata conversion from proprietary formats to standard formats and the custom NRGL format, full configurability of input and output format, (system description of the structure of input and output formats like XSLT for XML), the system should include settings for conversion to and from general and well-known formats, and the possibility for editing these descriptions via the user web interface
- formatting of bibliographical metadata for various purposes, e.g., various search and document presentation outputs, separation of data content from the administration of its output appearance, potential editing of these descriptions via the user web interface
- an automatic data collection mechanism (harvesting), support for general standards based upon OAI-PHM (Open Archive Initiative Protocol for Metadata Harvesting) and support for bidirectional interoperability with other digital libraries, web harvesting (downloading data via the internet), web submissions (provision of data via the internet interface from authorized entities, including conversion of fulltext documents from other text and image formats), e-mail upload, support for searching and other digital libraries
- a system for indexing metadata, references and full text files and mapping these indexes onto metadata tags for rapid searching in the database, administering index definitions
- a system of record classification in the database based upon various criteria for later search (e.g., frequency of appearance, tag value, number of records shown, number of documents requested etc.)
- checking entry data against the existing database with settable criteria, e.g., in order to avoid duplication in data being stored
- administrative tools for managing, configuring and administering the entire system, including managing operations on the central database and their timing with the option for manual intervention (e.g., shifting time-intensive tasks underway to a later period), access management for the system on the basis of roles (RBAC – Role Based Access Control)
- mass access via pre-formatted data (XML) to the central database, including controls on data consistency (bulk upload)
- a user web interface for searching, presenting and providing documents with support for various superstructure functionalities such as personal user boxes for document storage (similar to shopping carts), the setting

of various notifications concerning the appearance of documents with defined keywords, user discussion groups, user communication tools (a noticeboard, user commentary and evaluation of documents etc.)

- a search tool with a web interface enabling searching by word and phrase, including complex inquiries using Boolean criteria, structured search result presentation (classified according to type of document), recommendation of alternative searches when the original search fails to return results

Based upon this general specification of functionality main requirements for the software initially gathered, a more specific division of the criteria into 13 categories was made for purposes of selecting software for the digital repository and deciding the tender:

1. **Content management** – tools and procedures for storing content in digital libraries and controlling the process of storage and versioning
2. **Content generation** – import and export of content, support for different document formats
3. **Metadata** - Metadata support for various formats is important for indexing, storage, access and protection of content
4. **Search support** – the entire range of searching and browsing functions, metadata searching, fulltext search, hierarchical browsing etc.
5. **Access control and protection of privacy** – access control and protection of privacy includes password administration, user accounts with access rights including the ability to obtain a lost password etc.
6. **Support for output and query** – this criterion relates to the possibility of evaluating the use of digital libraries and uncovering user behavioral patterns to improve services, as well as recording user activity for billing purposes
7. **Sustainability, data safeguarding** – safeguarding of metadata, consistency and integrity of the database, backup, support and potentially data migration
8. **Interoperability** – option of bidirectional integration with other distributed systems at the metadata level, search as well as acquisition and provision of documents, OAI-PHM, Z39.50 support
9. **User interface** – this category focuses on support for multiple languages and the ability to customize the user interface to meet the needs of different users and different implementations
10. **Support for standards** – standards are important for sharing and long-term preservation of digital content. At issue is an entire spectrum of standards from metadata to interoperability to document storage formats

11. **Automation tools** – this category concerns tools for automated content acquisition, harvesting, metadata generation, maintenance etc.
12. **Support services** – an important aspect of all software systems. Many good digital library systems are Open Source, making it all the more important. Documentation, the helpdesk, collection of improvement requests, any discussion forums etc. are all important.
13. **Hardware and operating software** – hardware requirements for the data storage center, system operation and backup, ensuring access, supporting operational and database systems etc.

Criterion groups 1 to 9 and 11 to 12 are given in detail in the evaluation table along with weights for individual criterion groups and, for some, especially important individual criteria. Group 10 – Support of standards has been broken down in the table into relevant criterion groups and Group 13 – Hardware and Operating Software is located outside the table proper, with space provided for entries. The table is not shown due to lack of space.

There is good coverage for digital libraries in Open Source software, with open source code. When certain conditions are complied with, this openness means technical and legal accessibility to the source code for users resulting in economically advantageous acquisition of quality software. Open Source naturally does not mean free-of-charge software. Here, too, it is necessary to seek a certain level of stability of development, access to support services and software development etc. A secondary criterion in the choice of NRGL software, then, was the evaluation and comparison of the best-known programs selected for the area.

CDS Invenio, DSpace, Eprints, Fedora and Greenstone systems were evaluated and compared. These systems were selected more or less intuitively, based upon their broad use and recognition globally and in the Czech Republic. The comparison was focused on the basic characteristics of the systems selected and carried out at a relatively high level of abstraction for informational support of the selection process.

CDS Invenio

CDS Invenio[104] (formerly CDSWare), is a highly modular system for storing and making large numbers of library records available (Vesely et al. 2004). Created and further developed in Switzerland *European Organization for Nuclear Research* (abbrev. CERN). Development began in 1993, primarily for the internal needs of this renowned scientific institution, as an institutional repository. Starting in 2000, it began to include multimedia support, OAI-PHM etc. and was circulated under the GNU GPL license outside CERN, as well. Now a mature solution, developed in collaboration with *Ecole Polytechnique*

[104] CERN Document Server

Federale de Lausanne (abbrev. EPFL), for general document management, institutional repositories and extensive library systems. It has been localized into 18 languages, including Czech, Russian and others. It allows the definition of custom metadata schemata. The internal naming convention is based on MARC 21 and supports the OAI-PMH standard.

The CDS Invenio systemis highly complex, sophisticated and flexible. It can be set up in many different ways for the most varied purposes at the cost, naturally, of its relatively complex setup, maintenance and operation. It possesses the most formalized support for installation and operation of the systems included in this evaluation, provided under a paid contract. Eprints provides a similar level of service and support. Aside from the system itself, there is also localization of some services available in Czech and Slovak.

CDS Invenio consists of many more or less independent modules with precisely defined functionality, which together form a highly functional and efficient unit capable of further development and expansion using a defined interface for other modules from other manufacturers. A brief description follows of these which clarifies other characteristics of the system.

BibCheck is a tool for the administrator and content manager which automates various tests for metadata quality such as integrity rules, the format and extent of fields etc.

BibClassify automatically extracts keywords from a document based on the frequency of the terms contained in the dictionary (thesaurus).

BibConvert allows metadata conversion from any structured format to other formats. It works similarly to the XSLT processor. The input and output formats are fully configurable, descriptions of default formats are included in the product.

BibEdit allows editing of metadata through a web interface.

BibFormat enables the formatting of bibliographic data in many ways. It separates administrative content from output formatting and can work in the background in real-time and pre-format often-used outputs.

BibHarvest is a harvester compatible with OAI-PMH which allows metadata to be obtained from other repositories supporting OAI-PMH and storage management. It is created directly within the database and contains an administrative module which enables administration of the repository independent of main data administration. The repository may be completely or partially open for harvesting within the range of the OAI-PMH protocol. Data is provided in unprocessed form, with individual tags labeled according to MARC21 conventions.

BibIndex indexes metadata links (references) and text files. Creates two types of indexes - word and phrase. Is capable of defining a number of logical

indexes (author, title etc.). Indexes consist of two parts - a forward and reverse index. Such indexing allows for extremely rapid scanning.

BibMatch filters XML input files using the contents of the database and seeks records according to various criteria, e.g., to restrict duplicate records.

BibRank allows various sorting and ranking criteria to be set for later use in searching, for example, by frequency of occurrence of certain words, the number of document downloads etc. BibRank is independent of the module BibIndex.

BibSched is a task timer which serves as a central module for the entire system and enables other modules to gain access to the database in a controlled manner. It thus eliminates access conflicts to the database and ensures consistency in updating. It enables administrators to monitor and manage the task queue.

BibUpload enables new data to be saved to the database. Input must consist of a well formatted XML file which takes into account the metadata scheme chosen. In the majority of cases, this input data is the result of working with the BibConvert module.

ElmSubmit is a gateway for the automatic recording of documents from open sources via e-mail. It normally gives preference to the provision of files and harvesting from the internet.

MiscUtil is a set of tools for ad hoc use by other modules.

WebAccess is a module providing user access to the system. It uses the RBAC (Role-Based Access Control) technique in which users are divided into several groups according to their roles and permissions within the system. Each group is allowed only certain operations within the system.

WebAlert provides alerts to users of system events such as the appearance of a new document which meets certain criteria. These criteria correspond to typical user queries. Alerts can be sent to users by e-mail, or saved in their shopping carts.

WebBasket allows users to store documents in a personal shopping basket, just as users of online shopping outlets store goods. A single user may have multiple baskets. The basket may be personal or public. Public baskets allow documents to be shared by a group of users.

WebComment is a group tool for evaluating documents and sharing document commentaries among readers. Integrated into the WebBasket, WebGroup, WebMessage and WebComment tools is a tool center for group communication within the CDS Invenio system.

WebHelp is the global documentation of the CDS Invenio system for users and administrators. Documentation specific to individual modules is included in the pertinent module.

WebMessage enables communication between (possibly anonymous) users of the system via websites, for example, invitations to discussion groups etc.

WebSearch is a module used to query the system for specific search terms or phrases present in the database. It allows two types of queries: term search or phrase search. It also enables complex Boolean inquiries, searching using regular expressions, combined searching in both metadata and in the files themselves. Users may browse records using indexes, as well. If the query is unsuccessful, the system offers alternatives and search help. Metadata in collections is directly accessible for browsing similarly as with the popular *Web Directories* concept. Orthogonal document views are organized into virtual collections and documents may be classified using multiple criteria. This flexible organization allows the creation of a highly transparent navigation scheme for users.

WebSession is a relation and user manager which allows differentiating users and personalizing items in the user interface such as shopping baskets etc.

WebStat is a configurable system that enables information and statistics to be collected concerning the overall system and its use.

WebStyle is a library of modules defining the design and appearance of the CDS Invenio system.

WebSubmit is a comprehensive system that allows authorized users to store individual documents in the system. The content provision system includes control mechanisms, conversion of various document formats and the extraction of bibliographical data.

More information may be found at http://cdsware.cern.ch/invenio.

DSpace

DSpace, like most Open Source digital library systems, hails from the university environment. Its development began in the MIT library and at Hewlett-Packard, with the first version dated 2002. A total of 500 other universities and organizations in 60 countries currently collaborate on the project. Development takes place under the aegis of the nonprofit organization DSpace Foundation in Massachusetts. Development and bug reporting is taken care of via SourceForge. Clear developmental rules have been designated, with each member of the community able to contribute new functionality.

DSpace is software for creating digital repositories of different types of documents for varied types of organizations. It currently supports the Dublin Core metadata format and determines persistent identifiers using the CNRI Handle System,[105] with support for OAI-PMH 2.0 as data provider. It also

[105] http://sourceforge.net/

supports the OpenURL standard and allows data export to simple XML or the METS format.

Data is stored in the system via digital objects with unique identifiers containing both data and metadata. The system also includes structures for further logical division of records into collections and libraries. An object may belong to more than one logical structure. DSpace supports the creation of user accounts, grouping of users into groups, assignment of access rights and a basic user authentication system.

The system may be searched either using metadata or directly within the files themselves. In addition to pure text files, other formats such as PDF a DOC are also available.

A relatively quality internet user interface has been created using CSS, making it relatively easy to modify. A relatively recent arrival is the Manakin system, another version of the DSpace user interface in which the interface may be modified using templates. The user interface enables all basic functions including browsing and searching records for users and saving records for content administrators.

As regards the input of content, it is possible to define working procedures, including several levels of approval before content is actually included in the database.

The DSpace system has a relatively broad, active community of users and can be set up fairly easily but its functionality is limited and modifications require programming. It also specifically requires the UNIX operating system and a number of operating components.

More information may be found at: http://www.dspace.org.

EPrints

Eprints is a comprehensive system with many installations around the world, once again stemming from a university environment, in this case the *University of Southampton* in Great Britain and is available under the GNU license. The entire system is based upon web technologies. The primary goal is to create an institutional repository for many types of document, extending from ordinary literature to scientific and research work, as well as other various types of documents such as multimedia documents primarily focused on research information. Although the system is highly configurable and a general repository may be created as part of the implementation, this may be complicated by the system's emphasis on research. Paid services may be utilized during installation and operation, as well as during training, project implementation management, technical support etc., all in English.

As concerns standards, the EPrints system supports OAI-PMH. Metadata are represented using a special internal format. The system allows XML data to

be imported, along with data from some external sources (PubMed and others). Export may be done in many formats including XML, RSS, DublinCore, METS.

The internal architecture does not use digital objects but rather Items. The Item encapsulates both metadata and the file itself. Multiple repositories may be created as part of a single installation for different organizations, as well, or multiple separate collections for a single organization with individual appearances and structures.

Eprints supports the administration of user accounts but the assignment of rights is less sophisticated. The system distinguishes three types of user accounts – administrator, editor and ordinary user. Even ordinary users may enter data into the system with a high degree of detail, especially at the metadata level.

The EPrints system, like DSpace, features a comprehensive user interface based upon web technologies. The interface can be configured relatively well. A portion of the interface is written using static HTML generated by the system for items entered but is not designed for a large number of records to be entered daily.

The interface may be used to search metadata and data. In addition to text files, the system also indexes other common formats such as PDF. Content may be browsed using a logical tree structure, as well, with the default structure identical to that of the *Library of Congress* of the USA, but amenable to modification. The interface also allows registration of new users, information on news to do with the system, system administration etc. The entire system is quite comprehensive. Item entry is sophisticated but relatively complicated and demanding of time. The system is primarily designed for publishing researchers who enter a small number of units with extensive information content in each. This feature may complicate the creation of general repositories.

More information may be found at http://www.eprints.org.

Fedora

Fedora[106] also originated in a university environment at *Cornell University and the University of Virginia* (Fedora Commons, 2008). It began as a research project in 1997 and 1998. In 2001, both universities merged their work and received further development funding from the Mellon Foundation to develop a universal digital library on the basis of XML and web services. In 2007, both universities founded the Fedora Commons organization, which currently administers development of the joint system.

[106] http://www.handle.net/

The Fedora system is a digital repository which enables the storage and maintenance of virtually any digital object. Repository services are nevertheless offered only in the form of a core which provides for storage and administration of digital objects plus programming libraries of functions and calling services enabling developers to operate the repository. Aside from the basic GNOME interface, Fedora has no user interface of its own. It must be created separately for each installation of the system. In spite of the fact that the Fedora system provides a high quality repository built upon fundamental theoretical knowledge in the field of digital library construction, the necessity for creating a user interface in setting the system up prevents its use in rapid deployment projects and those lacking adequate developmental capacity. There are, of course, quality third party tools available for creating a user interface such as eSciDoc – a set of services using the SOA architecture, the RepoMMan development tool etc.

The fundamental architectural element of the system is, as with other systems, the digital object. But here, the object is not just a stored record; the system differentiates three types of objects:

- a Data Object, containing metadata and data and potentially a unique persistent identifier, pointing to data in another repository. The system makes versioning possible – the saving of different versions of the same document,
- a Behavior Definition Object which describes the behavior of services bound to specific files (e.g., displaying data of a particular type). These services are described in the WDSL language, ensuring relative platform independence for other types of objects,
- a Behavior Mechanism Object which may be, e.g., software enabling data display. These services may theoretically be written in any programming language.

The Fedora system supports a range of standards – OAI-PMH, export to METS format and the special FOXML internal format, with metadata descriptions stored in Dublin Core format.

Although the core of the Fedora system is very advanced, it cannot currently be rated as a complete library system. It is a quality foundation which must be complemented at relatively high cost and effort. Even during operation, the platform independence of the system requires more overhead and places more demands on hardware than other systems.

More information may be found at http://fedora-commons.org/.

Greenstone

The Greenstone system had a more exotic origins than its colleagues, but it is once again a university product. It has been developed since 1995 at the *University of Waikato* in New Zealand and its development is also supported by UNESCO. It is available under the GNU GPL. In spite of its exotic origins, the system has been partially localized into Czech, among other languages. The system supports many digital object formats.

As concerns standards, it supports the Dublin Core metadata scheme, the OAI-PMH protocol and is the only system reviewed to support the Z39.50 protocol for bidirectional catalog search of other libraries. This support is nevertheless not standardly available; its use requires recompiling.

The system's basic unit of storage is the document. Documents are stored in preprepared collections whose document types have been defined. The collections are then organized into libraries. The system must contain program support, i.e., plug-ins, for working with individual document types. This architecture places greater demands on the setup of the system than is true with other, more complete systems. A unique indicator is also generated in a special format when documents are saved into predefined collections in the internal XML format.

The user interface is separated from the actual repository as an independent application. The user interface and repository communicate using an internal internet protocol which may be replaced by the CORBA protocol, allowing an external application to be connected to the system. The user interface may be set up for each collection differently and is based upon web technologies. It may be used to browse collections and search within them. The collections are browsed using a classification of related documents into common folders which may be hierarchically connected, thus creating a logical browsing structure.

The system contains user administration and user access rights but support for this is not particularly sophisticated. It is confined to granting access rights to administrators and content administrators but does not address user rights.

The system is certainly viable. Services are also available, provided by *DL Consulting* in New Zealand, but its use is relatively complex and is limited to organizations in which data will not be entered by users other than the collection administrator.

More information may be found at http://www.greenstone.org.

The following table indicates approximate numerical scores for individual systems for the evaluation criteria selected as shown above, using a scale from 1 to 10:

	CDS Invenio	DSpace	Eprints	Fedora	Greenstone
Content management	10	8	8	5	9
Content acquisition	10	9	10	8	9
Metadata	6	5	5	7	5
Search support	8	6	5	4	7
Access management	8	6	8	4	5
Support for outputs	6	6	4	6	10
Sustainability, data protection	8	8	7	10	6
Interoperability	6	6	6	7	8
User interface	8	8	7	5	7
Support for standards	9	9	9	9	10
Automation tools	9	7	5	5	8
Support, services	9	7	8	6	6
Total:	97	85	82	76	90

If the systems introduced are evaluated strictly according to their characteristics without regard for their suitability for NRGL, the Greenstone and CDS Invenio systems come out on top, closely followed by DSpace. Fedora is not a ready-made system and completing it is costly and labour-intensive. Eprints is a relatively single-purpose system for smaller daily growth involving thoroughly described research documents.

In the Czech Republic, access to services and language localization make CDS Invenio a good choice. The system is rich in functionality and flexibility, the trade-off being its relative complexity of implementation and operation.

The software selected, a description of its functionality and a brief overview of system administration is given in the succeeding chapter.

REFERENCES:

VESELÝ, Martin – BARON, Thomas – Le MEUR, Jean-Yves – ŠIMKO, Tibor. *CERN Document Server: Document Management System for Grey Literature in a Networked Environment*. Publishing Research Quarterly, Spring 2004.

Introduction to Fedora, Fedora Commons 2008, Inc. Available from www: http://www.fedora.info.

9. NRGL Pilot Implementation

Jindřich Dolanský and Petra Pejšová

The aim of the pilot implementation of NRGL is to create an effective automated data collection (with both metadata and digital documents), to archive the data long-term and allow users pleasant, easy access. To reach this goal, software requirements were specified for a digital repository and associated services described in greater detail in the foregoing chapter. These served as a basis for a public tender. The NRGL software functionality requirements were defined in such a way as to include both requirements for pilot implementation of the system and a basis for selecting a modern, well-supported technology with good developmental prospects. To prepare for selecting software for NRGL, an analysis of selected Open Source software for digital libraries was undertaken (treated in the second section of the foregoing chapter). A public tender was subsequently held during which a software solution best suiting the requirements was selected based upon the CDS Invenio system for digital repositories and indexing, and the FAST ESP[107] search system for the user interface. This solution attained the highest ranking on the given criteria, as well as reflecting the results of an analysis which was further developed to offer a sophisticated solution. The solution led to a user interface being constructed for the digital repository which enabled easy, enjoyable 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.

This chapter describes the CDS Invenio system for digital repositories in detail from an administrative standpoint, since a preinstalled local installation for grey literature collections is being offered to cooperating organizations for installation on their own hardware. The ESP FAST indexing and search system is described here only briefly from the user's standpoint and not that of the administrator, because only a certified administrator may operate it and the services being provided by an external firm under the project. The chapter concludes with a further description of initial assumptions for the pilot implementation dealing with persistent identifiers and digital document formats.

CDS Invenio, described in the preceding chapter, was also selected for the NRGL repository because it is Open Source software. It may be freely installed, used and modified, which allows it to be set up for storing grey literature with subsequent distribution to cooperating organizations.

[107] http://www.incad.cz/jnp/cz/www/produkty/fast/-content-www-produkty-fast-fast_esp.html

The CDS Invenio system was installed during the first phase in a basic version distributed outside CERN. The virtualization application VirtualBox, which may be installed on virtually all recognize platforms, was used.[108] CDS Invenio was installed using VirtualBox over a Debian Linux operating system. This solution enables easy distribution of the preinstalled version of NRGL to cooperating institutions.

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 for further searches for related information. These functions are described further in this chapter.

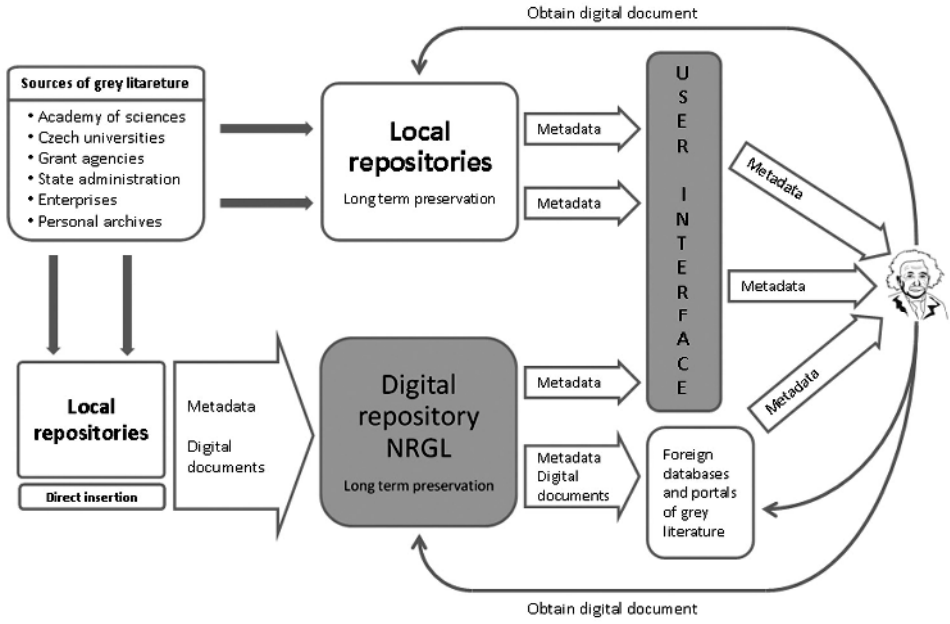

Figure 6: Scheme of the NRGL software solution

An advantage of the FAST ESP system is its ability to create an integrating search platform over a repository of academic resources containing grey literature. This makes it possible to create a comfortable central search interface even in the pilot phase of the project which is capable of allowing

[108] The VirtualBox virtual environment supports both Windows (primarily XP) and many Linux distributions (e.g. Debian, Ubuntu, CentOS, Fedora etc.). Since version 1.6, VirtualBox since version 1.6 has also supported Mac OS X and Solaris.

users access to data from both the NRGL digital repository and the distributed grey literature repositories in a single interactive environment. This trend represents a shift from building collections to tools which provide search and access to data, as Joachim Schöpfel noted in the second chapter in the section entitled From the Collection to the Open Portal.

Individual activities and digital repositories collaborating in CDS Invenio and the FAST ESP indexing and search system are depicted in the graphic.

CDS Invenio as the NRGL digital repository

The CDS Invenio system allows input and data acquisition, archival and management of stored data and offers wide range of access management. These functions will be described gradually from the point of view of the CDS Invenio system and its functionality.[109] CDS Invenio format issues will be described further and it will be demonstrated that although there are no advances restrictions on the software solution itself, restrictions must be defined for future administration of the archive. Finally, the NRGL software components which enable CDS Invenio to be online will be described.

CDS Invenio basic functions

This section will describe the basic functions of the CDS Invenio system, with an aim at presenting the variability achieved in the system thanks to its modular architecture. This architecture permits administration of a repository holding a large volume of data and despite the complexity of services, maintains speed.

Input/data acquisition

The CDS Invenio system enables data to be acquired by various means, as is illustrated in the figure depicting the CDS Invenio workflow diagram. Because of its modular architecture, CDS Invenio carries out each method of data acquisition with a corresponding specific module. There are three methods for data input:

1. direct input into the system,
2. obtaining bibliographic metadata by harvesting other OAI repositories,
3. obtaining metadata via e-mail.

[109] This text will not go into details which are available in the documentation itself.

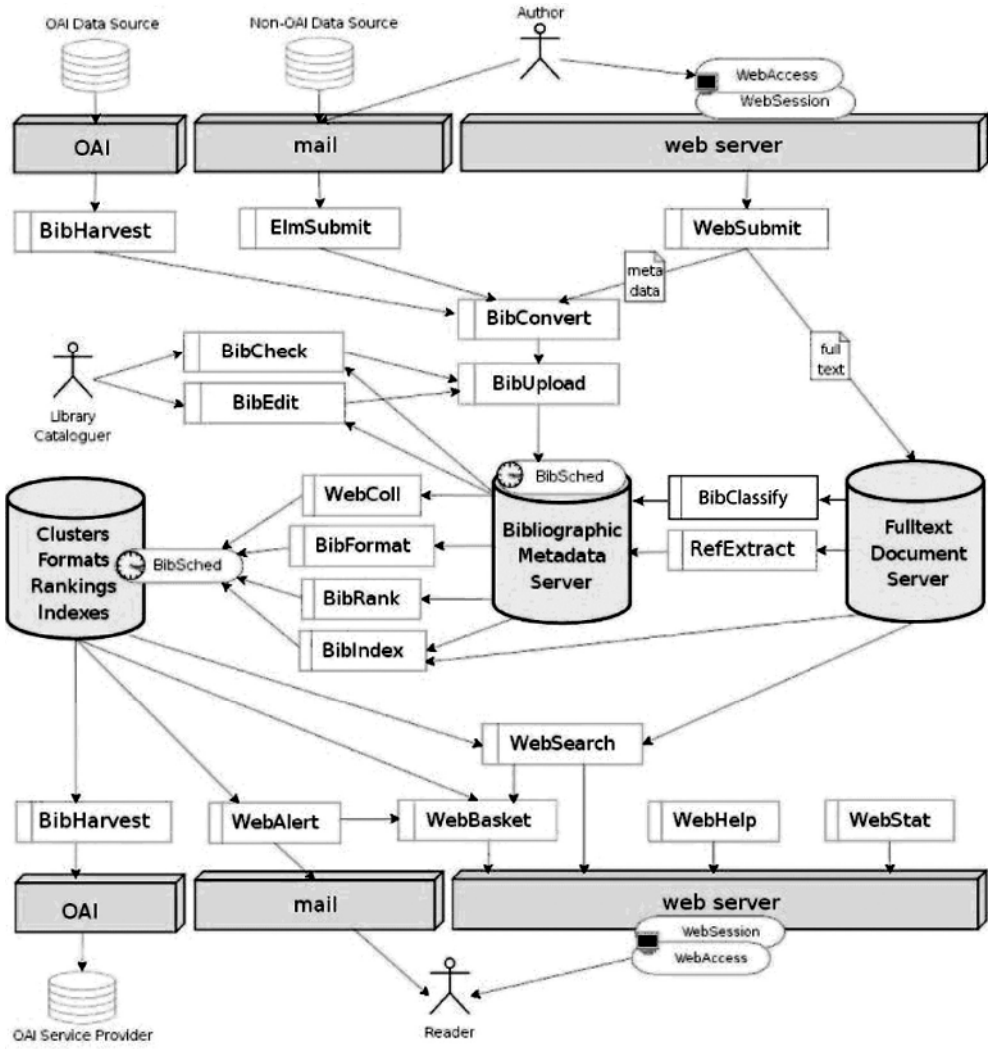

Figure 7: Workflow diagram (Pepe, 2005, pg. 6)

System input – the WebSubmit module

The WebSubmit module presents a set of functions used for storing metadata or the documents themselves. Configuration of the functions is done using the web interface. This interface provides intuitive configuration tools:

1. WebSubmit element definition

   WebSummit elements which will be used subsequently are defined in advance. These elements essentially constitute a MARC field into which the content will later be stored. These are actually classic form elements such as text entry, text fields and drop-down menus. A

special case is an entry for file upload, making it possible, e.g., to store digital documents.

2. Action definitions

   Actions to be carried out are defined but not implemented. These include, e.g., entering records, modifying records, entering files etc.

3. Definition of document types and categories

   Document types are defined (e.g., conference materials, reports, final theses), along with their inclusion into categories (posters, presentations, annual reports, research reports, degree theses).

4. Implementation of actions

   An action is selected for implementation for each document type. An interface will be created for each action comprised of defined elements and generate an internet form. After being filled in and sent, this is then processed via a series of functions providing for, among other things, record entry, creation of an ID for the record, renaming and moving files to the appropriate directory, extracting data from the form and transforming it into MARCXML format, assigning the transformed file to the upload queue, providing notification that the transaction has been successful, sending an e-mail to the uploader and backing up the original data.

5. Controls on input

   Data input by the uploader[110] may be checked using functions written in JavaScript. Verification can be done of whether the uploader has entered calendar data or names in the correct format.

Harvesting[111] metadata – the BibHarvest module

The BibHarvest module is a tool for defining and harvesting from other OAI[112] repositories. Configuration is done using a web interface, with the process itself run from the command line. In the web interface, parameters are defined for harvesting individual repositories. These include the name of the repository, the start point of the harvesting process (either from start-of-file or a particular record), the conversion method for the data and entry of the converted data into the database.

Actual harvesting of the OAI repository defined in the web interface is then initiated from the command line. If the pertinent oaiharvest command

[110] We employ the term “uploader” in this text to mean the person who is authorized to enter and modify metadata and upload digital documents to the CDS Invenio system.
[111] Harvesting is a commonly used term for data gathering.
[112] OAI-PMH Open Access Initiative - Protocol for Metadata Harvesting.

contains a periodicity parameter, it is assigned to the periodic task queue administered by the planning daemon BibSched.

A further parameter is the harvested format of the metadata. Metadata in OAI repositories is normally available in two or more formats. The Dublin Core format is mandatory under the OAI-PMH protocol. Also available will be at least one other format, which may either be based on standard formats such as MARCXML or be custom designed for the institution in question, for instance the EVSKP-MS format for eVSKP.

Harvesting metadata using e-mail – the ElmSubmit module

The builder of the system, CERN, is currently developing other means of data acquisition for CDS Invenio. This consists of the ElmSubmit module which allows documents to be entered into the database by e-mail. This functionality is also shown in the workflow diagram.

Data archival and administration

Fig. 7 shows that data are stored by type on three distinct servers and different mechanisms applied to their content:

1. Fulltext document server

   The fulltext document server contains text (or audio, visual or multimedia) files entered together with bibliographic metadata. It is advantageous to choose a concrete format for each file type to permit better future administration. See below.

2. Bibliographical metadata server

   The server for bibliographical metadata is connected with the fulltext server via reference numbers identifying the appropriate document for the record. Before data acquired using the OAI-PMH protocol are stored on the server, there may be a need for conversion using the BibConvert module based on transformation files describing modifications required for individual metadata elements. This module may also be used to convert metadata from an older system.

3. Format, classification method and indexing server

   If formatting, classification or indexing are carried out on metadata or fulltext, content originates on the third server.

Data administration

Applicable modules may be applied to individual types of data. As Fig. 7 shows:

1. BibClassify module for automatic classification

   The BibClassify module is used to automatically extract keywords according to the controlled vocabulary and heading system. BibClassify works with headings system in two formats. The first format is a heading system consisting of a simple list of keywords and a second heading system making use of the RDF SKOS taxonomy enabling the extraction of so-called concepts with a richer, more complex structure. Simply put, the BibClassify module presumes keywords will occur in fulltext which are contained in the heading system and subsequently orders these by frequency. RefExtract functionality may also be applied to fulltext to extract the list of references contained in the document. This functionality can only be used for PDF format.

2. BibConvert module for data conversion

   Metadata undergoes a basic modification process upon conversion to the CDS Invenio internal format by means of the BibConvert module. This module is either called by other modules (e.g. WebSubmit), or may be run manually, in, e.g., converting an XML file as part of a prescribed transformation. For working with metadata in a database, the uploader has two modules at his disposal. The first module, BibEdit, allows a web interface to be used to directly access a particular list which can be modified, added to or have individual fields erased as required. The second module, BibCheck, allows further editing and modifications from the command line.

3. BibIndex module for creating lists

   Other tools may be applied to metadata. The basic tool is the BibIndex module, which may be used to index a selected metadata field. The BibIndex module allows for configuration of so-called word files that define the metadata field for the word indexes. The metadata field for author, for example, is used to create a list of all authors. These word indexes are then used during searching. For example, the index for the author field is assigned MARC fields 100 and 700. Records may also be displayed in various metadata formats which may be defined using the BibFormat web interface. Metadata may also be classified based upon defined criteria and the classification used further for searching. An example would be searching similar lists based upon similar word content. Classification criteria are defined by the BibRank module.

It is advantageous to run some of the CDS Invenio modules noted periodically, for example, reindexation, formatting, reclassification after acquiring new records and regular harvesting of repositories or regular reestablishment of the local repository (or the subsets of its so-called sets) for harvesting by other institutions. Periodic running of these processes is provided for by the BibSched planning daemon, which is used not only to plan tasks for other modules but also to set various priorities.

Creation and configuration of collections is taken care of by the WebSearch module. The WebSearch module is used to define the search interface for searching collections and searching itself (indexes etc.). In the case of NRGL, collections are structured into two bilevel trees based upon document type and institution.

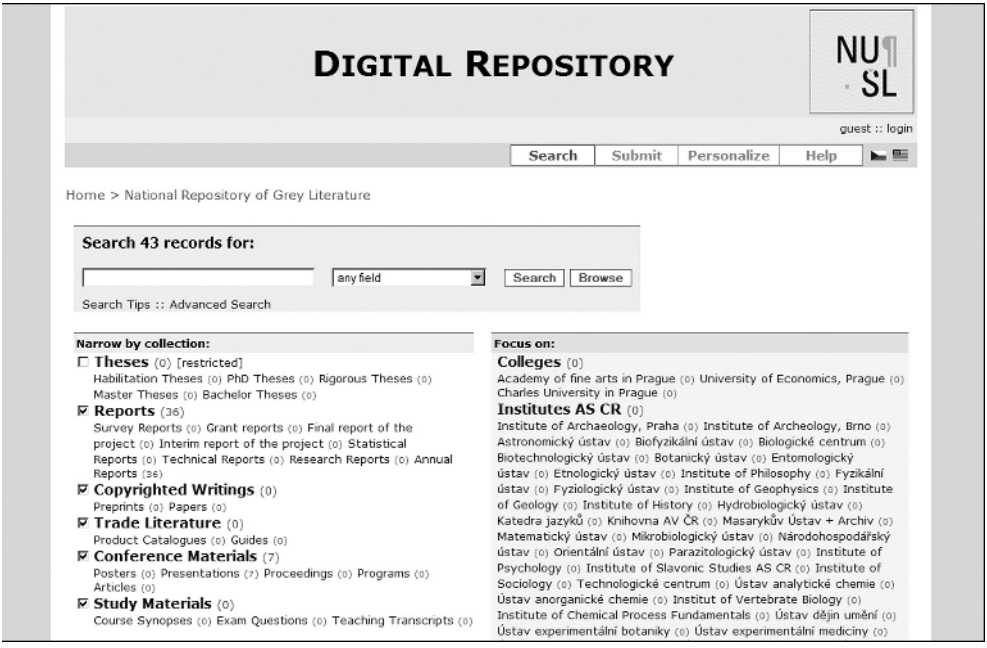

Figure 8: The NRGL digital repository is accessible at http://invenio.ntkcz.cz/

Collections for entering documents in the NRGL digital repository

A) Based upon document type

Collections - groups of document types, e.g., conference materials, final theses, reports etc.

Subcollections - individual types of documents, such as annual reports, final reports, dissertations, theses etc.

B) By institution

Collections - grouping by institution type, e.g., institutes of the AS CR, universities etc.

Subcollections - individual institutions, e.g., the Geological Institute of the AS CR, the Geophysical Institute of the AS CR, the University of Economics etc.

Another important module is the module OaiArchive, which allows the CDS Invenio archive to serve as an OAI data provider, as well. Data obtained whether by direct entry, harvesting or otherwise, is processed using the OAI-PMH protocol in MARCXML format (or Dublin Core format). The CDS Invenio archive thus also plays the role of an OAI data provider to the harvester mediated by BibHarvest. The CDS Invenio system allows further definition of repository sets, which may be specified for individual collections. Within these individual collections, subsets may also be selected, identified by phrase conditions, e.g., on the basis of the values of a particular metadata field.

Document administration

Issues to do with the long-term storage and accessibility of documents over a period of decades are also included in NRGL. It cannot currently be determined in advance which formats will be available for display on software used in 50 or 100 years. For example, Text602 format was among the most popular text formats in the first half of the 1990s on Czech personal computers. Although only five years have elapsed since its final version was issued, many users would have problems opening and reading these documents without the appropriate converter.

Document access may be solved using emulation/virtualization of the operating system and editor used during the time the document was created or by migrating the document to a more suitable format. When migrating documents to newer formats, it is important to ensure the maximum possible authentication of the document, even if this means it is sometimes impossible to avoid certain changes in appearance. It is thus good to preserve the original document during migration in case of errors in the migration process or if migration is later repeated using a better conversion tool. Document formats used for long term storage may be divided into:

1) the preferred format appropriate for long-term storage in the repository, which is a prerequisite for long-term usability,
2) accepted formats which will require greater care in upcoming years but do not require direct migration to the preferred format,

3) unacceptable formats, i.e., formats which must be migrated to the preferred format or at least to an acceptable format during the repository storage process.

The issue of long-term storage and access to grey literature in NRGL has been resolved on several levels:

1) by the selection of preferred, accepted and unacceptable formats for documents in accordance with a resolution of the government of the Czech Republic dated 3/11/2008 No. 1338 on the designation of output data formats of static text, image and combined character in digital form (Resolution, 2008) and by the approach of the National Digital Archive (BERNAS[113]),
2) By choosing migration with preservation of the original document as the default strategy for ensuring document readability,
3) by developing methodological guidelines to provide for regular tests of accessibility for individual formats of documents saved in NRGL and a strategy of potential conversion to more suitable formats.

Access rights

Three groups of users of the CDS Invenio system may be distinguished with regard to repository access:

1. Administrator

   The administrator is responsible for the actual operation of the system, primarily configuration of individual modules and where appropriate, modification of source code. Also assigns access rights for these modules for uploaders, saves user accounts etc.

2. Applicant

   The uploader receives access rights primarily for entering, editing and approving digital documents.

3. User

The CDS Invenio system allows the level of openness of the system to be set for users. The user has the right to access metadata without limitation and all public digital documents. Depending upon setup, rights may allow access to some nonpublic digital documents by collection, e.g., for a set of given IP addresses. Within his own user account, the user may change the e-mail address and password. He also has the opportunity to review documents from selected collections etc.

[113] http://knihovna.nkp.cz/knihovna91/bernas.htm

Access is configured by the administrator using the WebAccess module via the internet interface. This module enables access rights to be defined for individual administration and user activities. Setting restrictions on user access rights to certain collections, e.g., for a set of IP addresses, also naturally includes working with configuration files and other scripts.

Other software

CDS Invenio uses the Apache Web server and MySQL database server. But it's good to start with the CDS Invenio system installation itself, which requires installation of the GNU Linux operating system Debian, Gentoo) or the VirtualBox virtual environment.

VirtualBox

Ordinarily, those interested in operating CDS Invenio must create a package of freely available source code which must then be installed and configured. NTK has prepared a default version of NRGL for cooperating organizations which requires only the installation of VirtualBox.[114] This application creates a virtual environment on the local server which may contain one or more virtual machines. CDS Invenio will be run over the Linux Debian operating system on one of these virtual machines.

Upon installing VirtualBox, neither Debian nor CDS Invenio need be installed. The entire virtual machine is imported into VirtualBox, including the default configuration of CDS Invenio. After setting the network parameters (IP, hostname etc.) the virtual machine may be started. The aim of installing VirtualBox instead of CDS Invenio directly is to avoid dependence on the local server operating system, because VirtualBox can be installed on all well-known platforms.

In addition, a virtual machine with CDS Invenio may be exported to *Open Virtualization Format* (abbrev. OVF), as well as to another virtual environment other than VirtualBox. The oft-used virtual environment VMWare ESXi has also been tested.

The Apache Web Server

As noted above, CDS Invenio makes use of the standard Apache Web server with open code. The secure communications between the user and the server, CDS Invenio requires certified HTTPS protocol. This is used to secure login into the user account and registration of user accounts. During installation, CDS Invenio can generate a so-called self-signed certificate.

[114] http://www.virtualbox.org/wiki/Downloads

When CDS Invenio is published to an external internet network, it is of course essential to ensure a trustworthy public certificate. The virtual project and the CDS Invenio system make use of the TERENA server certificate for organizations belonging to the Czech national research network CESNET2, provided by Comodo CA Ltd., one of the largest global certificate authorities.

MySQL database server

Installation of CDS Invenio also requires installation of the MySQL database server, during which process all necessary tables and other database tools are created. The MySQL database is easily implemented, freeware and optimized for speed.

Python

Because CDS Invenio is programmed in Python, a Python interpreter must also be installed. In addition, it is also necessary to add the mod_python module to the Apache server in order to integrate the Python interpreter into the Web server. With this module in place, web applications may be effectively written in Python to work with the database.

NRGL central search interface

A central NRGL server for easy searching of grey literature in the Czech Republic has been available since 2009. Institutes of the AS CR are gradually connecting to the interface, as are universities, grant agencies and others. The interface may be used for access to metadata and digital grey literature documents in the areas of research, science and education. The central search interface contains Web 2.0 functionality and provides data visualization and dynamic contextual navigation which adapts with more precise inquiries. All connections are dynamic and based upon the current query and enable work with document relations.

The navigation shows textual filters at screen left (see graphic below) including document type, person, keyword and language. Document type navigation is structured into two levels, e.g., conference materials are further divided into papers, collections and presentations. The navigation for persons includes authors of documents as well as other persons connected to the creation of the document, e.g., the thesis advisor for final theses. Because the source database provides terms in either Czech or English, the navigation keywords include both languages. Keyword and person navigation show a default of 4 terms and a maximum of 25 terms. Language navigation offers a choice of all languages available for digital documents via the interface. The corresponding number of records is always listed in the navigation for each individual item, e.g., Document Type – Conference

materials - Papers (3,573). The number of records given in parentheses may be seen for individual source databases connected in the toolbar at top. The figure shows that NTK contains 9,129 records, the VSE source 19,062 records, etc. The image shows another search aid, a timeline. The timeline may be used to select a particular year or time period. Selection using the timeline always requires activation using the "Use" button.

An inquiry may be gradually built by a selection of search criteria from the navigation or by entering text into the search window. It is once again possible to make the search more precise using the navigation, which changes after each refinement and offers pertinent variants. Upon entering the query, the so-called used filters appears at left above the navigation. This makes possible refinement of the inquiry and may be eliminated with a simple mouse click.

Basic information concerning the document is shown in listing. Detail information, including document accessibility, appears by placing the cursor on the icon left to the title. The title contains an active link to the source database directly to the record in question. Document availability is controlled by database rules.

Searching in the central interface may be done at **www.nusl.cz**.

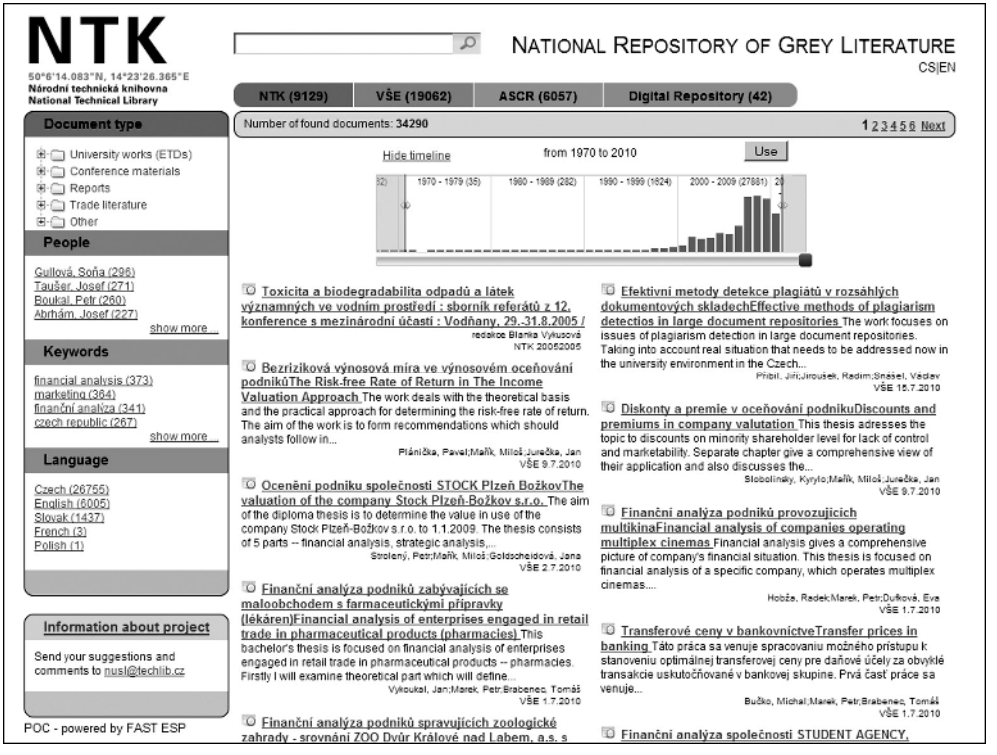

Figure 9: NRGL central search interface

Starting requirements for NRGL pilot implementation

Accepted digital document formats

Another important criterion for a digital archive is the means by which it will work with digital documents and the possibilities offered. Digital documents may be text, audio or video files. Because digital document formats change frequently, thought should be given when building an archive to the maintenance of digital sources in current formats. Otherwise, the archive could quickly become unreadable. In other words, it is necessary to select the broadest possible range of formats for which conversion tools exist enabling regular migration to the current version. Typical formats exist for all types of digital objects, e.g., PDF and DOC for text documents etc. The possibilities offered by the archive itself for the formats in question must also be taken into account.

In principle, CDS Invenio places no restrictions on digital document format. This is regardless of whether it be a text document, graphic or multimedia file.

To ease administration and digital document manipulation, e.g., during migration to a current format, it is good to restrict the number of acceptable formats. For the initial implementation of NRGL, formats were restricted to PDF text files, JPG graphics and MP3 and AVI multimedia files.

CDS Invenio preserves the original formats and does not transform them to other standard formats. It is therefore not possible to convert other formats to the standard format in an automated way up on import without help from another application.

Persistent identifiers

The primary purpose of digital archives is to archive digital information and make it accessible. This persistence of access to digital documents is a basic requirement which must be met by a persistent identifier. The problem of persistent identifiers made its appearance in the digitization of documents placed on the internet. Traditional identifiers such as LCCN 1898, ISBN 1970, ISSN 1975 – ISMN and ISAN, assigned to physical documents, function well in a non-digital environment. Persistent identifiers should play a similar role in the digital sphere. But in the digital environment, the relationship between the identifier and the document is more complex.

The easiest means of identification, using a URL link, has proven to be the least reliable Internet addresses change often. Links may be broken, files may be relocated (because of website reorganization, web domain changes, ownership changes etc.) and as a result the need has arisen to ensure digital

document references are permanent. This is the primary function of a persistent identifier. In addition, persistent identifiers should retain their permanence without regard to the permanence of the document to which they point. It is therefore important that the source pointed to by a persistent identifier is not moved or liquidated without updating the information concerning it in the persistent identification registry.

Additional requirements for persistent identifiers

In addition to persistence per se, additional requirements may be made of the identifier closely tied to the virtual environment in which the digital documents appear.

1. The identifier must be unique. Uniqueness of the identifier means that the relationship between the identifier and the digital document to which it points is unique in both directions. Uniqueness is provided for by the central authority of the system adopted. Identifiers must be unique within the digital resources of an organization but may also be unique worldwide. Global uniqueness may be attained if the system used is in operation around the world.
2. The identifier may be able to cover various levels of documents (e.g., books, book chapters, book images etc.). This capability is labeled granularity.
3. The identifier may be able to cover various versions of the work.
4. The syntax of persistent identifiers may reflect the needs of the given institution, e.g., substrings in the identifier may point to various subcollections.
5. Decentralized assignment is provided by an authority such as the administrator and other institutions with delegated powers (Bellini, 2007, pg. 8).
6. Interoperability with other systems of persistent identifiers.

Examples of persistent identifiers

There are several persistent identifier systems in use around the world, such as ARK, DOI, NBN and PURL. These systems were selected by various institutions for particular purposes and thus meet various requirements. The corresponding persistent identifiers therefore get their functionality and characteristics from the systems in which they function. For example, the Handle system with the DOI persistent identifier is used in the Czech Republic why the Library of the AS CR. The URN:NBN persistent identifier is used primarily by national libraries in, e.g., Italy, Germany, Switzerland, Hungary, Sweden, Holland etc.

The URN:NBN persistent identifier was chosen for metadata records in Ertl, with the National Library working on implementation in the Czech Republic. Until such time as URN:NBN is initiated in the Czech Republic, local identifiers are assigned within NRGL which will be easily expanded when URN:NBN is fully online. The syntax for NRGL local identifiers is nusl-number (= NRGL number), e.g., nusl-31. This value is entered into the help field in internal CDS Invenio MARC21 format. CDS Invenio does not require zeros before digits at a particular position count within the number. The local identifier is then used to create a permanent link, which is the sole variable. The entire permanent link looks as follows: http://www.nusl.cz/ntk/nusl-number, e.g., http://www.nusl.cz/ntk/nusl-31 and is entered into the dc:identifier in NRGL format. Subsequently, a proxy server providing the functionality of a permanent reference to the identifier was created.

This chapter has clarified description of the software chosen for the NRGL pilot implementation. Attention was devoted chiefly to the CDS Invenio system used for long-term grey literature archival and data access. Standards required for system functioning and data administration were noted – data formats and persistent identifiers. The ESP FAST system allows search of both the NRGL repository and other repositories of academic resources containing grey literature in the Czech Republic. At the time of writing, the NTK catalogue, VSE final thesis database and ASEP AV CR system were connected. Further resources will constantly be added to the interface and partnership networks with cooperating institutions will be expanded. An audit in the DRAMBORA system was carried out to ensure the long-term proper functioning of the NRGL digital repository.

REFERENCES:

PEPE, A. – BARON, T. – GRACCO, M. – Le MEUR, J., Y. – ROBINSON, N. – ŠIMKO, T. – VESELÝ, M. CERN *Document Server Software: the integrated digital library*. CERN, Geneva, Switzerland, 2005.

BELLINI, E. – CIRINNA, C. – LUNGHI, M. – DAMIANI, E. – FUGAZZA, C. *Persistent Identifiers distributed system for Cultural Heritage digital objevte*. Dept., University of Milan Via Bramante, Crema, Italy, 2007.

BERNAS, Jiří. *Národní digitální archiv*. Knihovna, 2009, issue. 20, no. 1, pp. 22 – 29. Available from www: http://knihovna.nkp.cz/knihovna91/bernas.htm.

ELECTRONICS SOURCES:

Usnesení [online]. Usnesení vlády České republiky ze dne 3. listopadu 2008 č. 1338. 2008. Vláda České republiky. Retrieved August 27, 2010 from www: http://racek.vlada.cz/usneseni/usneseni_webtest.nsf/0/22AE7ADE48C55BF8C1257553002951A5/$FILE/1338%20uv081103.1338.pdf.

10. An Audit of the National Repository of Grey Literature using the DRAMBORA tool

Petr Karlach

An audit of NRGL as a trustworthy digital repository using the tool and methodology called the *Digital Repository Audit Method Based on Risk Assessment* (abbrev. DRAMBORA) was planned and performed at the end of 2009 as an integral part of creating a digital repository of grey literature in the NTK.

Methodologically, the audit stemmed from experience with the audit of the digital university repository at Charles University, which uses the same methodology and tool, and with the audit of a digital repository at the NL CR, which was processed using an offline version of the DRAMBORA tool.

As was true of preliminary planning for NRGL itself, the preliminary preparation phase was of high importance for this audit. In the preparation phase, the audit staff acquainted itself with up-to-date trends and methods for planning, creating and certifying trustworthy digital repositories. Furthermore, the staff was introduced to the on-line version of the DRAMBORA tool[115] and audit methodology – this audit is run internally, without participation by any external specialized auditing firm.

The DRAMBORA tool was used during the preparation phase in addition to the audit phase. The preparation phase consisted of acquiring all relevant information and documents on the status of the repository, its description, standards, procedures, staff, material, budget etc. (Donnelly et al., 2009). This information served as input data for the preparation phase of the audit and was entered into the DRAMBORA database in the section "Before the Assessment". Here, the repository was described, the scope of relevant areas (Functional Classes) of the audit were defined and the repository staff, including a detailed description of individual team members and their roles. The definitions of staff roles were especially important since at the subsequent risk identification stage, it was necessary to relate risks to respective roles. Even during the preparation phase, a great contribution may be made to an audited repository. It helps the staff to see the repository from a global vantage point, to map and accumulate the most important descriptive data about the repository and to point to possible deficiencies and defects offering the opportunity for problems to be remedied and missing materials to be completed.

[115] For more information see http://www.repositoryaudit.eu.

The audit was run using the portion of the DRAMBORA tool called the Assessment Centre (Donnelly et al., 2009). Here, the repository mandate was defined: its mission, raison d'être, founder etc. Other repositories were also identified influencing its activity, both external (e.g., legislative) and internal (e.g., organization, content type restrictions etc.). The audit continued by defining repository goals, activities and means used to achieve goals. The goals were to be quantifiable strategic aims, for instance the target number of records in the database for a given time span, and were to relate to the appropriate functional classes. Activities are mapped according to the roles of the repository team and related to the means utilized. The final phase of the audit the identifying risks currently endangering the operation, the stability and existence of the repository or which could do so in the future.

For inspiration and possible direct help, the DRAMBORA tool contains a number of links to supporting documents and a range of practical examples of completed entries, whether in the preparation phase or the audit phase. In the area of risk identification, pre-defined risks can be directly used and modified or unique risks may be formulated. Risks are mapped onto staff roles and functional classes and may be related to limitations, activities and sources. To define risks, two factors are considered: the probability of their occurrence and their presumed impact on repository operation. The tool thus defines the rate of their severity. Furthermore, strategies of risk handling are defined – how to prevent risks and deal with the consequences, as well as a schedule for checking anti-risk measures.

When all relevant information is entered, the Reporting Centre function creates output reports on risks identified in the repository, their severity, interconnection and solutions for them. There are two types of output reports available in PDF format and one in HTML format. Other descriptive information entered cannot be easily exported; however, with some effort, it can be copied from the snapshot of the updated audit state. After the audit is completed, the following logical step is to create an action plan aimed at eliminating or reducing the impact of risks identified, including the schedule of checkpoints set by the audit. Thus, auditing becomes a periodical process throughout the existence of the repository. For originally identified risks, a solution strategy is drawn up, the risks are repeatedly evaluated at checkpoints, and some risks may be eliminated. The audit is then repeated under the new conditions and the modified risks are evaluated once again.

Practical work with the online DRAMBORA tool presented one initial problem – after inputting access data, it was found that the tool does not support languages from the code page of ISO-8859-2/windows-1250, i.e., not Czech, resulting in the degradation of data already introduced and problems with access rights. Surprisingly, the tool has been written using the ISO-8859-1 code page, instead of common UNICODE UTF-8. The authors of the tool promised to eliminate the problem; however, the solution is not likely to be

quick as the problem is contained in the application and also the database setup where data are entered.

It was necessary to establish a new database for the audit and decide whether English or Czech without diacritics should be used. English was chosen for a number of reasons, primarily because using Czech without diacritics is unnatural; secondly, both the tool and methodology are available only in English; and, thirdly, the audit results will be presented to an international public. There is an offline version of the DRAMBORA tool, consisting of simple forms in Microsoft Office DOC and XLS formats. This version would allow full use of Czech; however, it offers practically no ease, no interconnection of information and no support when entering and evaluating data compared to the online version. Since the tool was selected just for these features, the offline version was considered useless.

After solving these initial difficulties, work with the tool seemed to be relatively straightforward except for some minor faults which could be overcome. For instance, when risk factors for the repository and their relations were to be entered, only a portion of the data filled in advance was available. The solution was to regress one step and repeat the action. The main problem of the online version proved to be the absence of back-up copies, ideally with the possibility of export if a damaged or lost database had to be re-created. At some stages of the work, input information was suddenly unavailable and the database issued only error messages; nevertheless, the database authors reacted quickly enough for the most part in solving the situation.

To sum up, the tool itself is quite easy and intuitive to use. A much bigger problem consists in a quality preparation phase. It's not enough just to prepare the available documents and basic information; a map should be drawn up of the entire audit and the mutual relations between all of its parts. Although the demands made by this phase were not underestimated by any means, data input sometimes required regressing one step and to filling in missing connections. One of the chief contributions made by the tool is to reveal connections and the project proved very well prepared in this respect. The NRGL documentation, description of the project and related documents were at a good level and represented a good basis for the audit.

It is necessary to point out that this preparation phase will be a great contribution to the audit of other repositories in the future, if the repository was not constructed and planned by employing a strict methodology such as PLATTER and does not implicitly contain the requirements of the audit. This method is scarcely suitable for any repository, since it is very consuming of time and finances and the first useful results are produced only after a long wait. A detailed map of the repository using the DRAMBORA tool helps to get an overview of what is ready and what is not, which important documents, procedures, means and measures are missing and where the most severe risks lie endangering the existence and success of the repository, how individual

parts of the repository are related, which documentation is available etc. The data in the audit database may be linked externally using the appropriate URL, or other documents may be attached. Thus, the database may be considered a center of important information on the audited repository.

In addition to the repository mapped and the relevant environment, the producers of the methodology and the tool consider the most important output to be the analysis of identifiable risks endangering the repository, its quality, readiness, reputation and position in the eyes of both specialists and ordinary users. NRGL is being audited in the pilot project phase, i.e., early enough to use the audit conclusions in the full operational phase. NRGL was analyzed in all functional classes but only the most severe and most obvious risks were recorded according to Pareto's rule of 80/20 – 20% of risks are responsible for 80% of the danger. In the following paragraphs, the risks and conclusions identified in the first stage of analysis are shown to illustrate audit outputs. It must be borne in mind that these risks only serve as a static snapshot illustrating NRGL's situation in the past.

Risks

As noted above, the audit focused on the most severe and most obvious risks in the pilot phase of the project and its subsequent transition to full operation. These risks are mainly related to the description of repository activities and procedures, to the state and development of the repository staff, to project funding, and to hardware and software sources including their back-up and the relevant NRGL environment. Sixteen basic risks were identified – their order is not relevant since the numbering originated in the DRAMBORA tool.

Risk Number 1: Loss of Staff Members

This risk considers the fact that the NRGL team is relatively small, recently created and as such can hardly cover all related activities or solve the problem of substitutability if some team members are absent or leave. Thus, NRGL activity is directly endangered by any absences. Ideally, the repository team should be expanded in such a way that knowledge could be shared and substitutability could be increased; however, this is beyond the project budget. Another suitable solution would be to complete and broaden the descriptions of knowledge and activities of individual team members, possibly also in electronic form as is used in Wikipedia, the free encyclopedia. Thus, any team member performing substitution could be quickly acquainted with his new activity.

This risk is related to risks nos. 6, 12, 15 and optionally to risk no. 16.

Risk elimination carried out by June 30th, 2010: The project team is gradually expanding, along with substitutability. All team members continuously document their activity and present it in regular meetings. Working documents are then saved in the non-public part of NRGL wiki.

Risk Number 2: Pilot Project Termination

One of the NRGL mandates follows from the project of the Ministry of Culture of the Czech Republic for the period of 2008-2011 – The Digital Library of Grey Literature – Functional Model and Pilot Realization. This project will come to a conclusion in 2011, which may cause weakening of the NRGL mandate and possible financial problems. This challenge may be met by applying for participation in another similar project or looking for other financial sources to continue and develop the NRGL project.

This risk is related to risks nos. 4 and 6, partially also 9.

Risk elimination carried out by June 30th, 2010: The functions of long-term archival and data access, including the acquisition of new data are and will be covered by the NTK budget after 2011. For further development, another funding program will be located.

Risk Number 3: Disaster Recovery

There is a guideline for backup of the NRGL database but no plan exists for how to proceed in the case of recovery after a potential failure or damage to the database and who shall be responsible for individual steps. This risk may lead to prolongation of the time needed to recover or other problems. Such a plan must be prepared including the selection of responsible staff and connection to the backup plan, which also needs to be finalized and embodied in the NRGL wiki noted under Risk no. 1.

This risk is indirectly related to risk no. 5.

Risk elimination carried out by June 30th, 2010: The backup methodology and recovery plan in the case of repository failure are being worked on simultaneously. This document is currently being certified by the NTK.

Risk Number 4: Voluntary Partner Network

The network of partner organizations supplying material to NRGL is based upon volunteer participation; no law mandates its creation, all is exclusively based on the willingness of employees of partner institutions. This risk may cause difficulties in supplying material to NRGL. A possible solution consists in finding motivating factors for partner organizations (such as the benefits of a centralized repository, the international aspect of the project etc.), and in

encouraging the promotion of NRGL targeted to partner organizations and the academic public. Last but not least, motivation similar to publishing in specialized journals would be helpful at a political level – some kind of positive evaluation for organizations supplying their grey literature production to NRGL, possibly with an impact on their funding.

This risk is related to risks nos. 2, 6 and 9.

Risk elimination carried out by June 30th, 2010: Partners are offered long-term archival and local installation of the CDS Invenio system, preinstalled for building a grey literature collection. Guidelines and instructions are gradually produced that provide information about how to create and maintain grey literature collections in local repositories. Workshops for partner institutions are held on a regular basis. During 2010, NRGL is being promoted in specialist journals, at conferences, invited lectures and personal meetings with representatives of grey literature producers.

Risk Number 5: Backup Storage Tapes

Backup tapes containing copies of the NRGL database are stored in the same room as the hardware repository. If there is a fire, terrorist attack or other serious disaster, the database would be destroyed together with its backup. At the present time, under pilot operation, when the database only contains minimal records, this risk is negligible; however, in the full operational phase, an alternative storage location must be found and this fact must be specified in the finalized backup plan.

This risk is indirectly related to risk no. 3.

Risk elimination carried out by June 30th, 2010: see risk no. 3.

Risk Number 6: Financial Shortfall

Present budget and funding sources cover the project in its existing form but do not assure NRGL development in any relevant areas. No institution, project or company showing zero development and growth has the potential for long-term existence. The solution is the same as that for risk no. 2 – timely application for participation in a similar project or a search for other financial sources to support the continuation and development of the NRGL project.

This risk is related to the majority of other risks mentioned here.

Risk elimination carried out by June 30th, 2010: Other project funding is continuously assured, for instance, the license of the ESP FAST system and the plan of hardware renewal.

Risk Number 7: Budget for Services

Services supporting the ESP FAST software system, provided by an external supplier, are under contract and covered by the budget until the end of 2010. When the budgets are prepared for the NTK and the NGL for 2011, this item should not be neglected and discussions about continuing cooperation with the external supplier should be initiated in time.

This risk is related to risks nos. 2 and 6.

Risk elimination carried out by June 30th, 2010: Continuous cooperation with the external supplier of the ESP FAST system is now secured by contract until the end of 2013.

Risk Number 8: ESP FAST Trial Version

NTK is running the trial version of the ESP FAST software system based on a contract with an external supplier until the end of 2011. This software supports the NRGL user interface. During 2010, discussions about the purchase, operation and support of the full version should be initiated so that it may be used for routine operation of NRGL. The purchase of licenses and their support must be included in the appropriate budget.

This risk is related to risks nos. 6 and 7.

Risk elimination carried out by June 30th, 2010: This risk was eliminated by purchasing the full version.

Risk Number 9: Weak Mandate

The NRGL mandate is based on a time-limited project of the Ministry of Culture of the Czech Republic and on a brief note in the NTK status. This fact may negatively influence the reputation and position of NRGL in the eyes of the academic public and partner organizations at home and abroad. This problem should be solved at a political level, for example by gaining formal support from the NL CR, and AS CR, by promotion in the appropriate ministries etc.

This risk is related to risks nos. 2, 4 and partially 6.

Risk elimination carried out by June 30th, 2010: The topic of grey literature was included in the proposal for the concept of library development until the end of 2014. There is a continuous effort to resolve this risk by discussions with NTK management, responsible bodies and civil service units.

Risk Number 10: No Input Policy

There is no formal procedure and methodology for handling documents acquired, processing them, determining which checkpoints must be cleared before saving in the database, who is responsible for individual steps etc. This risk may result in a chaos when documents acquired are saved, resulting in insufficient processing and monitoring procedures. As a result, the accumulated data may lack consistency. This methodology must be drawn up, including nomination of responsible staff, and must be embodied in the electronic NRGL wiki, noted under Risk no. 1.

This risk is loosely related to risks nos. 11, 13 and 14.

Risk elimination carried out by June 30th, 2010: Acquisition methodologies and methodologies for the processing of grey literature documents are gradually being prepared. The role of the repository team content manager was established.

Risk Number 11: Document Formats

Although the typology of NRGL documents is well described, formats for documents to be saved in the NRGL database are not sufficiently defined. This is also true for formalized procedures involving potential migration of formats to newer versions, acquisition of new software needed to access these new versions etc. This risk may result in the inconsistency of saved documents and problems accessing them over a long time span.

A plan and described methodology should be prepared defining which document formats will be supported by NRGL and how their migration to upgraded versions will be done, including the software necessary to access the documents.

This risk is loosely related to risks 10, 13 and 14.

Risk elimination carried out by June 30th, 2010: This risk is being resolved by ordering an external analysis of this issue.

Risk Number 12: Software Administration

Administration of the NRGL repository software system – CDS Invenio – depends on a single individual within the NTK. In addition, this person is not a formal member of the repository team controlled by the NRGL manager but is a member of the IT department. If he is absent, there is no acceptably rapid means to substitute for him. This could seriously endanger NRGL operation.

A possible solution may be to hire another person and thus increase substitutability in software administration, or, alternatively, to hire an external company under contract. The latter seems more suitable but tends to be more financially demanding. This fact should be considered during budget preparation.

This risk is related to risks nos. 1 and 6 and loosely to risk no. 16.

Risk elimination carried out by June 30th, 2010: This risk was substantially reduced by placing the administrator directly beneath the NRGL manager, by training another employee and preparing instructions for the administrator.

Risk Number 13: Undocumented Policies

Although the documentation of the NRGL project is at a very good level, the audit identified several specific points at which there is insufficient documentation or none at all. The audit concentrated on the most substantial facts and risks in the pilot project phase. After completion of the audit, the documentation of all basic processes must be analyzed, areas lacking coverage must be identified and a timetable for completion of relevant documentation must be created, including the electronic NRGL wiki encyclopedia – see Risk no. 1.

This risk is related to Risks nos. 1, 10, 11, and 14.

Risk elimination carried out by June 30th, 2010: This risk is gradually being eliminated by creating documentation for procedures still lacking coverage.

Risk Number 14: Long-term Preservation Strategy Not Described

A long-term strategy for storing digital documents is one of the most important aspects in operating and developing a trustworthy digital repository.
This strategy globally considers the vitality of the digital repository regarding technology (such as migration of document formats), functioning of repository staff and the organization as such, funding of repository operation and development etc. The rudiments of this technology are contained in the documentation of the NRGL pilot project; however, this strategy must be prepared as a separate document.
This risk is related to risks nos. 10, 11 and 13.
Risk elimination carried out by June 30th, 2010: This risk is being resolved by ordering an external analysis as in the case of risk no. 11.

Risk Number 15: Insufficient Staff Skills

As novel technologies and methods for building and operating digital libraries are developed, it may happen that the knowledge and capabilities of the repository staff will not keep pace with the latest developments. This risk endangers the staff and is especially relevant in the case of a small, busy team with practically zero substitutability, where there is insufficient time for training and education. NRGL management should not underestimate this problem and try to provide space and the means for regular training and education of all staff members. Naturally, this should be taken into consideration when the number of team members and NRGL budget are planned.

This risk relates to risks nos. 1, 6, and 16.

Risk elimination carried out by June 30th, 2010: The project team is gradually expanding, along with the substitutability of its members. Staff is regularly educated by individual study as well as at specialized conferences and workshops.

Risk Number 16: Deliberate System Sabotage

The risk of deliberate system sabotage by dissatisfied employees or employees on notice, damage to the database, long-term disruption of data consistency and other factors can never be excluded or ignored. Similarly as with risk no. 15, such risks are especially relevant in the case of a small, busy team with practically no mutual substitutability and with one member formally belonging to another department; moreover, if no formal procedure exists for status of the database. NRGL management should try to stimulate positive motivation on the team and encourage its professional development. On the other hand, functional backup should be implemented, along with control mechanisms such as periodic control of data consistency in the NRGL database.

This risk is related to risks nos. 1, 6, 12, and 15.

Risk elimination carried out by June 30th, 2010: The risk has been eliminated by daily backup of data and the system and by creating the role of content manager on the team. In order to support team building, activities often overlap with substitutability thereby increasing.

After the audit had been finished, an action plan was created analyzing individual risks and eliminating or moderating their impact on NRGL activity. As already mentioned, the above list of risks represents a static snapshot illustrating a past state. Many of these risks have already been eliminated or have begun to be resolved within the scope of the action plan. The current

status is always noted for individual risks on the list. A further step consists in repeating the audit under changed conditions, identify novel or modified risks, preparing a new action plan etc. Thus, the audit is becoming an interactive process of repeated risk identification and solution, and will contribute to the credibility of NRGL.

REFERENCES:

DONNELLY, Martin – INNOCENTI, Perla – McHUGH, Andrew – RUUSALEPP, Ravio. *DRAMBORA Interactive User Guide*. Glasgow 2009. Available from www: http://www.repositoryaudit.eu/help/.

Conclusion

With changes in the manner of publication as the shift is made from traditional print to electronic publication, many documents are created which may be classified as grey literature. The production and dissemination of information is now much easier thanks to novel technologies which allow institutions whose primary focus is not publication to disseminate information on a large scale. These publications may then be preserved at all levels using digital repositories of grey literature. It is around this pivotal theme that this entire book has been written.

The first section gives a general definition of grey literature and indicates its importance in Europe and the Czech Republic, as well as offering a description of the status of grey literature in institutions. One overarching theme recurs here: an emphasis on collaboration, whether it be at the local, national or international level. The second portion of the book has a practical orientation and provides general guidance on building a digital repository targeting grey literature. An attempt is made to answer questions such as what is necessary to build a grey literature repository, what software is available, what kind of descriptive data should be employed and how should grey literature be collected, archive and disseminated in accordance with Czech law? The knowledge communicated in this section derives from experience in designing the National Repository of Grey Literature project.

This knowledge will be leveraged by a team at the National Technical Library to focus further on research into grey literature digital repositories, to support academic discussion on the issue and to broaden knowledge of grey literature while building the National Repository of Grey Literature. The emphasis will continue to be on collaboration at both the national and international level. Under the project, the team will attempt to provide access to further Open Access resources in the Czech Republic and to support the sharing of information sources in science, research and development.

Appendix 1 Grey Literature Databases in the Czech Republic

Producer	URL	Description
Academy of Sciences of the Czech Republic	http://www.avcr.cz	A website which publishes current academic information for the academic community and the public at large.
Association of Libraries of Czech Universities	http://www.akvs.cz	It defends and supports the interests of universities and their libraries in negotiating with educational, information technology and communications technology organizations.
University of Pardubice - Digital Library	http://dspace.upce.cz/	UPa Digital Library, based upon the DSpace system and primarily created for eVSKP (electronic thesis) storage and access.
Charles University in Prague - Digital Repository	http://repozitar.cuni.cz/	UK Digital Repository, built using the DigiTool system.
Masaryk University in Brno - Digital Library	http://dspace.muni.cz/	MU Digital Library – system DSpace.
Academy of Performing Arts in Prague - Institutional Repository	http://dspace.amu.cz/	AMU Institutional Repository, in service since 2008 (system DSpace).
Jan Evangelista Purkyně University in Ústí nad Labem - Digital Repository	http://edu.ujep.cz/main.php?idwebu=dspace	UJEP Digital Repository for electronic theses (system DSpace).
Technical University Ostrava - Digital Repository	http://dspace.vsb.cz/	VSB-TUO Digital Repository, built and administered by the VSB-TUO Central Library (system DSpace).
National Register	http://www.theses.cz	National Register of Masters Theses and Doctoral Dissertations (Czech Republic).

National Repository of Grey Literature	http://www.nusl.cz	Central search interface for the National Repository of Grey Literature, providing access to grey literature both from the NRGL repository and other sources.
Information Register of R&D results (RIV)	http://www.vyzkum.cz	The RIV section of IS R&D collects information on the results of research and development and research projects supported by public funds under the Act on the Promotion of Research and Development.

Appendix 2 Grey Literature Databases Abroad

Database	URL	Description
ANRT	http://www.anrtheses.com.fr/	In France, the ANRT association develops developed activities aimed at digitization as part of their Theses à la Carte service.
ArXiv.org	http://www.arxiv.org	Preprints (physics, mathematics, computer science, nonlinear science).
CEDA Repository	http://cedadocs.badc.rl.ac.uk/	The digital depository is focused on grey literature in the Earth monitoring and atmospheric sciences. It contains more than 600 digital documents.
CNR	http://www.cnr.it/sitocnr/home.html	Institutional Repository of the Italian National Council for Research (CNR).
DART-Europe	http://www.dart-europe.eu/	The electronic thesis warbled provides more than 130,000 fulltext research papers from 233 universities in 16 European countries. DART-Europe is a partner to research libraries and library consortiums for the European Working Group NDLTD.
DOAJ	http://www.doaj.org	Scientific and academic journals (fulltext).
DOE	http://www.energy.gov/	One of the largest digitization projects in America, targeting report collections.
DFG	http://gepris.dfg.de	German digitization projects funded by DFG mostly include primary cultural heritage sources, along with specialized publications (manuscripts, journals, etc.).
DiVA	http://www.diva-portal.org/	The portal provides access to 270,000 research publications and student research papers from 27 Swedish and Norwegian universities and institutions of higher education. 44% of its content is grey literature.

DRIVER	http://www.driver-community.eu/	The project goal is to create a European network of digital repositories offering sophisticated functional services to both researchers and general public. It extends existing institutional repositories and networks.
EKT	http://www.ekt.gr/en/index.html	EKT Athens administers the Hellenic Dissertations Database with links to more than 13,000 works in Greek universities.
ETH	http://www.ethz.ch/	The University Library website provides access to 2.1 million reports from other libraries and databases, along with access to research tools.
EthOS - portal	http://ethos.bl.uk/	The British Library digitizes works from British universities for the portal, which combines free access to electronic dissertations from open repositories and works digitized specially upon request.
Grey Literature LibGuide (University in Calgary)	http://libguides.ucalgary.ca/greylit	Compilations of grey literature documents in healthcare and health sciences.
GreyNet	http://www.greynet.org	GreyText Archive: an internal grey literature document archive.
JISC	http://www.jisc.ac.uk/digitisation	The British JISC digitizes the content of special collections, e.g., historical censuses and masters theses.
LARA	http://lara.inist.fr/	The platform allows access to the full text of French academic reports in the humanities and social sciences, and contains almost 200 reports.
NARCIS	http://www.narcis.info/index	The NARCIS portal, developed by the KNAW in Holland, providing users with information on research, programs, projects, researchers and research institutions, along with their profiles.

NDLTD	http://www.ndltd.org/	The international organization is dedicated to promoting the adoption, creation, use, dissemination and preservation of electronic theses, with access to almost 750,000 works.
NTIS	http://www.fedworld.gov	Science, technology and business.
NUMDAM	http://www.numdam.org/	A mathematical archive with 29 seminars from the period between 1948-2007.
OpenDOAR	http://www.opendoar.org	Directory of open academic repositories.
OpenSigle	http://opensigle.inist.fr	Free access to bibliographic records of gray literature in Europe.
OSTI	http://www.osti.gov/	Scientific portal for grey literature focusing on technical reports. Freely accessible documents via the online interface from DTIC, DOE, NASA, EPA etc.
ProQuest	Předplacená služba	Database of theses from North American universities.
PUMA	http://puma.isti.cnr.it/	System for accessing the network of institutional repositories in Italy. Currently, access to 40 institutional repositories through the user interface in Italian or English.
ROAR	http://roar.eprints.org	Open repositories, e-print courses, educational materials.
SCIRUS	http://www.scirus.com	Allows search of the content of academic journals, academic websites, preprints, patents and institutional repositories.
TEL	http://tel.archives-ouvertes.fr/	French National Repository of Theses, with more than 10,000 works from the period 2005-2010 and almost 2000 digitized works published before 1990.
TDX	http://www.tesisenxarxa.net/	Catalan Network of electronic thesis and dissertations.
Web 2.0 Ernest Barrer	http://www.netvibes.com/ernestobarrera#Grey_literature	Personal page with RSS focusing on sharing internal knowledge concerning biomedical grey literature.

Appendix 3

The question arises as to whether theses are archival material, and if so, under what conditions they should be discarded. According to the Copyright Act, bachelors, masters and dissertations are characterized as "academic work" created under university supervision in order to complete study obligations. They may also be considered as a part of the protocol concerning the student's final state examination, which is definitely included among archival materials (Grulich, 2005).

The answer to this question may be found in sample discard guidelines, which allocate discard codes to individual documents:

A – documents with permanent documentary or information value destined for permanent archival,

V – documents reassessed after the protection period is over re-ranked into category "A" or "S",

S – documents that do not require further archival due to their documentary or information value after their operational or administrative function has been accomplished; these may then be discarded.

The number following the discard code stands for the length of the protection period in years. After this time, the document is proposed either for transfer to the pertinent archive (discard code A), for discard (discard code S) or is reassessed and sent either for archival or discard (discard code V).

According to Ministry of Education Regulation No. 11834/57 dated July 16th, 1958, and No. 10324/60-L dated April 1st, 1960, masters and similar theses were considered to be archival material with a discard code of A10. The amended Regulation, No. 19 151/87-491 dated 1987 allocated masters theses with a discard code of V20; dissertation and habilitation theses with a discard code of A20. Masters theses should be reassessed according to the criterion of their contribution to science and research. However, this reassessment was not done in practice at universities. Therefore, this amended regulation was omitted from the list of valid regulations in the area of education, youth and sports issued after 2005.

The current practice of these is archival differs substantially among universities as can be documented by looking at the sample discard regulations given in Table 1. The only outstanding feature is the acknowledged importance of habilitation theses, which are mostly classified as A5 or A10, exceptionally also as A40 (VŠE). Bachelors and masters theses are commonly classified as V20, as recommended in the above Regulation no. 19 151/87-491 of 1987. This overview clearly shows the varied importance accorded to bachelors and masters theses at individual universities.

The diversity is very obvious in the following examples. At MU, only papers of students are given a discard code of A10. In contrast, at Palacký University in Olomouc (abbrev. UPOL), bachelors and masters theses are unusually classified as S10, which enables these theses to be discarded after ten years without any reassessment. At Mendel University of Agriculture and Forestry in Brno (abbrev. MELDELU), an extremely short archival time of only five years was adopted disregarding the type of the final thesis.

Final thesis	VŠE	1st Faculty of Medicine, Charles University	UTB	UPOL	MENDELU
Bachelors	V20		V20	S10	V5
Masters	V20	A10	V20	S10	V5
Dissertation	V20	A10	A5	A	A5
Habilitation	A40	A10	A5	A	A5

Overview of codes for discarding final theses according to guidelines of selected universities

Appendix 4

In December 2006, the ETD Working Group performed the first survey on the ETD access. It shows that electronic versions of theses had already been collected by 10 universities (at 6 of them this was done mandatorily and at 4, optionally) before the amended Act was put into force (Horová and Krkošková, 2007). The survey demonstrated high adoption of the EVSKP-MS metadata standard, which was used by 14 survey respondents in 2006. In a repeated survey one year later, this unified format was planned for use by 16 survey respondents. Others either employed their own formats or the MARC/UNIMARC format. As regards the obligatory collection of theses, in 2007 it was found that 21 universities organized collection of theses. Of these, one was in pilot operation and two others had only facultative collections. Only two universities – six faculties of Charles University and The Silesian University in Opava – admitted no aims at ETD collection. Obligatory collection definitely prevailed (Horová, 2007).

The last survey, performed in 2009, monitors thesis accessibility at 16 universities after running the theses.cz system (Vyčítalová, 2010). Responses revealed that licensing contracts are made less frequently than before (13 respondents made no contracts, 6 respondents did and 2 respondents planned to). Libraries often save ETDs in library catalogues, and the MARC format is therefore more preferred (25 % of respondents) for thesis description. By contrast, only 25 % of respondents opted for the EVSKP-MS format; a great decline compared to 2006. Presumably, most universities are able to provide data export even in the EVSKP-MS format if needed, since the required data for metadata transformation are saved in the system.

TU is the only university to carry out retrospective digitizing of theses, planned to extend back to 1979. Three universities (AMU, JAMU, VUT) also process non-text theses (films, audiovisual records, photos etc.). In the *Registry of Scientific Information* (abbrev. RIV), dissertations related to projects monitored by the governmental Research Council are filed by six universities. One university also files habilitation theses.

A great proportion of universities saves ETD data in the theses.cz system manually (38 %) or in batches (47 %) and only 15 % of respondents use automatic data harvesting through the OAI-PMH protocol.

Index to abbreviation

A
AKVŠ – Association of Libraries of Czech Universities
AMU – Academy of Performing Arts in Prague
ANRT – L'Association Nationale de la Recherche et de la Technologie
ANSI - American National Standards Institute
ASEP - Automated System of Publication Registration, database of AS CR publication activities
AS CR – Academy of Sciences of the Czech Republic
AS CR Library - Academy of Sciences Library of the Czech Republic

B
BMBF - Federal Ministry of Education and Research; (Bundesministerium für Bildung und Forschung)
Boston Spa - Boston Spa conferences collection
BRGM - Bureau de recherches géologiques et minières
BUT - Brno University of Technology

C
CDL - Czech Digital Library
CDS – Cern Document Server
CEDA - Centre for Environmental Data Archival
CERN - European Organization for Nuclear Research
CNR – Consiglio Nazionale delle Ricerche
CTU – Czech Technical University in Prague

D
DART-Europe - E-theses Portal
DCAM – Abstract Model
DCMI - Abstract Model
DFG – Deutche Forschungsgemainschaft
DIDL - Digital Item Declaration Language
DissOnline - Digitale Dissertationen im Internet
DiVA - Academic Archive On-line
DOC – text document Microsoft Word
DOE – Department of Energy
DRAMBORA - Digital Repository Audit Method Based on Risk Assessment
DRIVER - Digital Repository Infrastructure Vision for European Research
DSc. – title “Doctor of Science“ in the Czech Republic

E
EAGLE - European Association for Grey Literature Exploitation
EKT – National Documentation Centre in Athens
EPFL - Ecole Polytechnique Federale de Lausanne
ETD - Electronic Theses and Dissertations

ETD db - ETD database
ETD-MS – standard for description ETD
ETH - Die Eidgenössische Technische Hochschule Zürich
EthOS-portal – Electronic Theses Online Service
EVSKP-MS - metadata set for ETDs in the Czech Republic
ETD Working group - Working Group under the aegis of the Association of Libraries of Czech Universities

F
FIZ Karlsruhe - Fachinformationszentrum Karlsruhe
FRBR - Functional Requirements for Bibliographic Records
FRV – University Development Fund MEYS

G
GAAV – Grant Agency of the Academy of Sciences of the Czech Republic
GACR - Czech Science Foundation
GL - International Confernce on Grey Literature
GreyNet – Grey Literature Network Service

I
ICC - Institut Cartogràfic de Catalunya
IGA – Internal Grant Agency
IILS - International Interlibrary Loan Service
INIST (CNRS) - L'Institut de l'Information Scientifique et Technique (Centre National de la Recherche Scientifique)
ISO - International Standard Organization

J
JAMU - Janáček Academy of Music and Performing Arts in Brno
JISC - Joint Information Systems Committee

K
KNAW – Royal Netherlands Academy of Arts and Science (Koninklijke Nederlandse Akademie van Wetenschappen)
KSŠL - Cooperative System for Grey Literature

L
LANL – Los Alamos National Laboratory
LARA - Libre accès aux rapports scientifiques et techniques
LIBER - Ligue des Bibliothèques Européenes de Recherche
LIGRIA - Littérature Grise Administrative
LIS - Library and Information Science

M
MIT - Massachusetts Institute of Technology
MENDELU - Mendel University of Agriculture and Forestry in Brno
METS - Metadata Encoding and Transmission Standard
MEYS - Ministry of Education, Youth and Sports in the Czech Republic

MODS - Metadata Object Description Schema
MU - Masaryk University in Brno

N
NARCIS - National Academic Research and Collaborations Information System
NATO - North Atlantic Treaty Organization
NDK - The National Digital Library
NDLTD - Networked Digital Library of Theses and Dissertations
NIWI – Institute for Scientific Information Services
NL CR - National Library of the Czech Republic
NRGL – National Repository of Grey Literature
NTIS - National Technical Information Service
NTK – National Technical Library, until July 1st, 2009, it was called the State Technical Library (STK)
NUMDAM - Numérisation de documents anciens mathématiques
NRGL – National Repository of Grey Literature

O
OA – Open Access
OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
OAIS - Open Archival Information System
OCCL - Online Computer Library center
OECD - Organization for Economic Cooperation Development
OPAC - On-line Public Access Catalog
OpenDOAR - Directory of Open Access Repository
OpenSIGLE - System for Information on Grey Literature in Europe
OSTI - Office of Scientific a Technical Information
OVF - Open Virtualization Format
OU - University of Ostrava

P
PI – Persistent Identifier
PDF - Adobe Portable Document Format
PPSNC DLT - Poznáń Supercomputing and Networking Center Digital Libraries Team in Poznan
PSH – Polythematic Structured Subject Heading System
PUMA - Akademisches Publikationsmanagement

R
R&D – Research and Development Council
R&D IS - Research and Development and Innovation Information System of the Czech Republic
RBAC – Role Based Access control
RIV – Index of Research and Development of the Czech Republic
RDA - Resource Description and Access
RDF - Resource Description Framework
RSS - Really Simple Syndication

S
SGML - Standard Generalized Markup Language
SIGLE - System for Information on Grey Literature in Europe
SKIP - Association of Library and Information Professionals of the Czech Republic
SKOS - Simple Knowledge Organization System
STI - Scientific and Technical Information
STAG - Information System of Study Agenda
STFC - Scientific a Technology Facilities Council
SUDOC - Système Universitaire de Documentation
SURA - Southeastern Universities Research Association

T
TDKIV - Czech Library and Information Science Database
TDX – Tesis Doctorals en Xarxa
TEL – Theses en Ligne
theses.cz - National Register of Theses and Dissertations and a System for Seeking Evidence of Plagiarism in Theses
TIB - German National Library of Science and Technology, (Technische Informationsbibliothek Universitätsbibliothek Hannover)
TU Delft – Delft University of Technology
TUL - Technical University in Liberec

U
UJEP - Jan Evangelista Purkyně University in Ústí nad Labem
UNESCO - United Nations Educational, Scientific and Cultural Organization)
UPa - University of Pardubice
UTB - Tomas Bata University in Zlín
UWB - University of West Bohemia

V
VNTIC – Moscow Scientific and Technical Information Centre
Virginia Tech - Virginia Polytechnic Institute and State University
VISK - Public Information Services Library
VŠB-TUO - Technical University Ostrava
VŠE – University of Economics in Prague

W
www - World Wide Web

X
XML - eXtensible Markup Language

Index

Grey Literature Repositories

Petra Pejšová (ed.), Marcus Vaska, Joachim Schöpfel,
Iveta Fürstová, Radim Polčák, Jan Mach,
Bohdana Frantíková, Petr Karlach, Jindřich Dolanský

Publisher: Radim Bačuvčík - VeRBuM
(Přehradní 292, 763 14 Zlín 12, Czech Republic)
Publishing editor: Marek Adamík
Zlín, 2010
First edition. 156 p. Printing: 200 ks
Print: Kodiak print, s.r.o., Zlín

www.verbum.name
www.verbum.webnode.cz

ISBN 978-80-904273-6-5