

An International Journal on Grey Literature



Autumn 2018, Volume 14, Number 3

‘PROFILING GREY DATA FOR PUBLIC ACCESS’

GreyNet

The Grey Journal

An International Journal on Grey Literature

COLOPHON

Journal Editor:

Dr. Dominic Farace
GreyNet International,
Grey Literature Network Service
Netherlands
journal@greynet.org

Associate Editors:

Julia Gelfand
University of California, Irvine
United States

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

Dr. Joachim Schöpfel
Université Charles de Gaulle Lille 3
France

Prof. Gretta E. Siegel (Emeritus)
Portland State University
United States

Dr. Plato L. Smith
University of Florida
George A. Smathers Libraries
United States

Technical Editor:

Jerry Frantzen, TextRelease

About TGJ

The Grey Journal is a flagship journal for the international grey literature community. It crosses continents, disciplines, and sectors both public and private.

The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.



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

CIP

The Grey Journal (TGJ) : An international journal on grey literature / Dominic Farace (Journal Editor); Jerry Frantzen (Technical Editor) ; GreyNet International, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 14, Number 3, Autumn 2018. - TIB (DE), DANS-KNAW (NL), FEDLINK-Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist-CNRS (FR), ISTI-CNR (IT), KISTI (KR), NIS-IAEA (AT), NTK (CZ), and the University of Florida (USA) are Corporate Authors and Associate Members of GreyNet International.
ISSN 1574-1796 (Print)
ISSN 1574-180X (PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:

€240 institutional, including postage & handling

Contact Address:

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

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

About GreyNet

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

Full-Text License Agreement

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

© 2018 TextRelease

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

Contents 'PROFILING GREY DATA FOR PUBLIC ACCESS'

D4Humanities: Deposit of Dissertation Data in Social Sciences & Humanities – A Project in Digital Humanities	113
<i>Joachim Schöpfel and H����� Prost (France)</i>	
Public Access to the Dissertations in Russia	119
<i>Yuliya B. Balashova (Russian Federation)</i>	
Data on Grey Literature Containing Scientific and Technical Information	123
<i>Leonid P. Pavlov (Russian Federation)</i>	
Indexing grey multilingual literature in General Practice: Family Medicine in the Era of Semantic Web	125
<i>Marc Jamoulle [et al] (Belgium)</i>	
Apps & Codes: Making profiles for fluid publishing contents	143
<i>Flavia Cancedda and Luisa De Biagi (Italy)</i>	
Data from "Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication"	149
<i>Dominic Farace (Netherlands) and Joachim Schöpfel (France)</i>	
Data from "Leveraging Grey Literature: Capitalizing on Value and the Return on Investment"	151
<i>Dominic Farace and Jerry Frantzen (Netherlands); Stefania Biagioni and Carlo Carlesi (Italy)</i>	

Colophon	110
Editor's Note	112
On The Newsfront	
Best Conference Poster 2018	154
GL20 Call for Posters - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	155
GL20 Registration - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	156
Advertisements	
KISTI, Korea Institute of Science and Technology Information	122
EBSCO LISTA - Library, Information Science & Technology Abstracts with Full Text	142
Author Information	157
Notes for Contributors	159

REPORT ON A TRAINING MODULE DEVELOPED FOR LONG TAIL RESEARCH DATA

Editor's Note*

In 2017, GreyNet embarked on a Data Papers Project in an effort to comply with the FAIR Principles endorsed by data repositories and archives worldwide. Compliance with these principles demonstrate that data/sets are findable, accessible, interoperable, and reusable.

To this end, a standardized template was designed and implemented as a tool in enabling authors/researchers to create and publish data papers within the GreyNet community as well as in other communities of practice involved in the long tail of research data. This refers to data/sets that are typically small in size, heterogeneous in format, and cannot easily be standardized with regard to curation and data management <https://www.radar-projekt.org/display/RE/Glossar>.

The results of GreyNet's project formed the basis for a training module that by the end of October 2018 will have been delivered in four workshop settings held in the United States (March 2018), Italy (May 2018), The Netherlands (August 2018), and the Czech Republic. GreyNet's training module consists in a stepwise approach signaling the data paper as a grey literature document type, demonstrating its value as a use case, and a trusted tool in research and data sharing. This training module is offered both standalone or in collaboration with other open data initiatives.

Dominic Farace,
Journal Editor

* TGJ now endeavors to link the ORCID iD with the name of each author published in the journal.

D4Humanities: Deposit of Dissertation Data in Social Sciences & Humanities – A Project in Digital Humanities *

Joachim Schöpfel <https://orcid.org/0000-0002-4000-807X>
Hélène Prost <https://orcid.org/0000-0002-7982-2765> (France)

Abstract

Following our work on research data and electronic theses and dissertations since 2013, we are conducting a new research project between 2017 and 2018 called D4Humanities with three objectives – to develop the research data management and stewardship on our campus, to gain better insight into the nature of research data in social sciences and humanities and to produce empirical evidence on the development of dissertations. In particular, the project contains three components:

- 1. Qualitative survey on behaviours and knowledge in the field of research data with 50 scientists from the University of Lille Social Sciences and Humanities Department, with a special focus on the FAIR guiding principles of scientific data management and stewardship.*
- 2. The creation of a workflow for the submission of research data related to PhD dissertations (deposit, preservation and dissemination of data via the NAKALA service Huma-Num)*
- 3. Two conceptual studies on the definition and typology of research data in SSH and on the development of dissertations in the environment of e-Science and Open Science (content, format, structure, requirements).*

In the following, we present some preliminary results, in particular from the survey and from the conceptual studies, in order to enhance the understanding of research data in SSH and of the development of dissertations.

Acknowledgment: The project receives funding from the European Institute of Social Sciences and Humanities (MESHS Lille) and from the Regional Council (Conseil Régional Hauts-de-France). We would like to thank the D4Humanities project team for their contribution to the research underlying this paper, in particular Cécile Malleret, Eric Kergosien and Leslie Hyacinthe.

Keywords: PhD dissertations, research data, digital humanities, open access, open science, social sciences and humanities

Introduction

For more than ten years, one part of our professional and scientific work has been focused on PhD dissertations as one major document type of academic grey literature. We started with research on their production and findability (Paillassard et al. 2005) and then moved on to questions related to their accessibility, especially in the environment of electronic theses and dissertations (ETD), open access (OA) and institutional repositories (Schöpfel 2013, Schöpfel & Prost 2013, Schöpfel et al. 2015c). Three years ago, we began to study the research data produced by PhD students and submitted as complementary material together with the dissertation (Schöpfel et al. 2014), trying to establish the link between grey literature and e-Science in the field of ETDs. Our first questions were (are) operational: what could and should be done with this material, how can it be stored and preserved for a longer time, what are the conditions for sharing, publishing and reuse? However, these practical questions always included conceptual elements, about the definition and typology of data, about their identification and description, about their relationship with dissertations, and about the development of dissertations themselves and their potential for reuse with content mining tools. Because of the complexity of the field, we limited our research to the disciplines of social sciences and humanities (SSH).

In 2017 we launched a two-year project called *D4Humanities*¹ in order to transform our research work into operational service development on the campus and to enhance our knowledge of data and dissertation. The basic question is how to enable the exploration of research data in social sciences and humanities (textual or oral corpus, raw data, images...) with digital technologies (text and data mining,

* First published in the GL19 Conference Proceedings, February 2018.

¹ Deposit of Dissertation Data in Social Sciences and Humanities. A Project in Digital Humanities <https://d4h.meshs.fr/>

mapping, visualization ...) to convey a new meaning? The project *D4Humanities* is part of the Digital Humanities and a continuation of the recent research of the GERIICO laboratory and its partners at the University of Lille Humanities and Social Sciences (academic library, SSH graduate school, digitization centre ANRT...) with the objective of accelerating the research data management project in particular for PhD students and young researchers, and of fostering the preparation of an international research project.

We started our project in March 2017, and it will continue until fall 2018. So what we will do here is deliver some preliminary results on data behaviour and data management, including the development of a workflow for ETD related datasets, and first conceptual work on data and dissertations. This will be followed by an invitation to join our research consortium.

Data literacy (survey)

In 2015, we conducted a campus-wide survey at the University of Lille on research data management in social sciences and humanities. The survey received 270 responses, equivalent to 15% of all scientists, scholars, PhD students and administrative and technical staff; all disciplines were represented. The responses showed a wide variety of data, practice and usage; some differences seem related to job status and disciplines. Generally, 20-25% of the sample can be considered as pioneers in data management and sharing, and 25-30% are motivated; only 5-10% appear reluctant to make their data available (Schöpfel & Prost 2016).

On the basis of the results of this first survey, we prepared a small qualitative survey with academic "volunteers" on the Lille SSH campus, among researchers and PhD students from various disciplines. We wanted to gain more insight in personal research data management behaviour and data literacy, in particular those contributing to the compliance with the FAIR principles for data management (Wilkinson et al. 2016). The investigation is not over; for the moment, we have conducted 27 interviews with researchers from history, archaeology, literature and language studies, psychology and information sciences. First results and comments:

Interest and motivation: finding volunteers on the campus was not easy this time; obviously, for many colleagues RDM is not a "hot topic" to spend one hour or more in a semi-directive interview on data practice and literacy. At least, it does not appear as priority or relevant.

Funding agencies: one half of the volunteering respondents (14) has conducted or participated in one or more research projects funded by the European Commission (H2020 program) and/or the French National Research Agency (ANR programs). But only 10 have knowledge of requirements (such as of the H2020 program), guidelines or recommendations for RDM.

Privacy: 13 respondents use or produce personal data as defined by the French CNIL commission, or confidential data. 6 submitted a research protocol to the university's ethics committee.

Standards, description: 8 participants reported assigning codes to their data, 9 people have already drafted a data management plan, and 5 participants follow standards for describing their data.

Dissemination and sharing: data collection, analysis and storage are often carried out by the researcher him/herself or together with the research team. 16 participants agree to share their data with others, which means above all with other colleagues from the project team. 10 participants have already submitted their data to an online server, 2 others intend to do so; only one refuses for security reasons.

Need for advice: generally, the respondents need advice on querying databases, formatting and naming data; they seek advice on licensing and legal protection of sensitive data; they want to know more about the services offered by the deposit platforms. So far, they have been seeking advice on RDM not at the library but with people from the IT department (system security, storage) and from the ethics committee.

Need for data services: the services requested by the researchers relate mainly to the different aspects related to data storage: to know what data to store, under which formats, on which server, with which guarantees of duration and security. They want to encourage exchanges between researchers and information professionals.

So far, we have observed very large differences between disciplines and research domains, but also between research methods and tools in the same field. Some scientists have a long experience with

RDM and apply standard and transparent data procedure, even if they do not always call it RDM. This data literacy can mainly be explained by legal issues (privacy laws, especially in psychology, education, sociology, and projects in public health) or ethics rules, less (up to now) by requirements from funding agencies. However, application of standards in RDM remains exceptional, such as data publishing and sharing. We did not encounter significant reluctance or even opposition to RDM and data sharing, but rather ignorance or lack of interest.

Data workflow

Similar to other ETD projects² we are developing a local workflow for the deposit of research data by PhD students. The main characteristics of this workflow are:

- Data and dissertations are submitted on different servers,
- The local deposit is interconnected with existing infrastructures, in particular with the French SSH data platform NAKALA,
- Data and dissertations are stored and preserved on various platforms but linked via their metadata and identifiers.

Figure 1 shows the workflow and the separation of data and dissertations from the beginning on (deposit). The guiding principle was to provide an interface (with technical assistance) on our campus for the deposit of research data on the NAKALA platform of the national infrastructure for SSH communities. For a detailed description, see Schöpfel et al. (2017b).

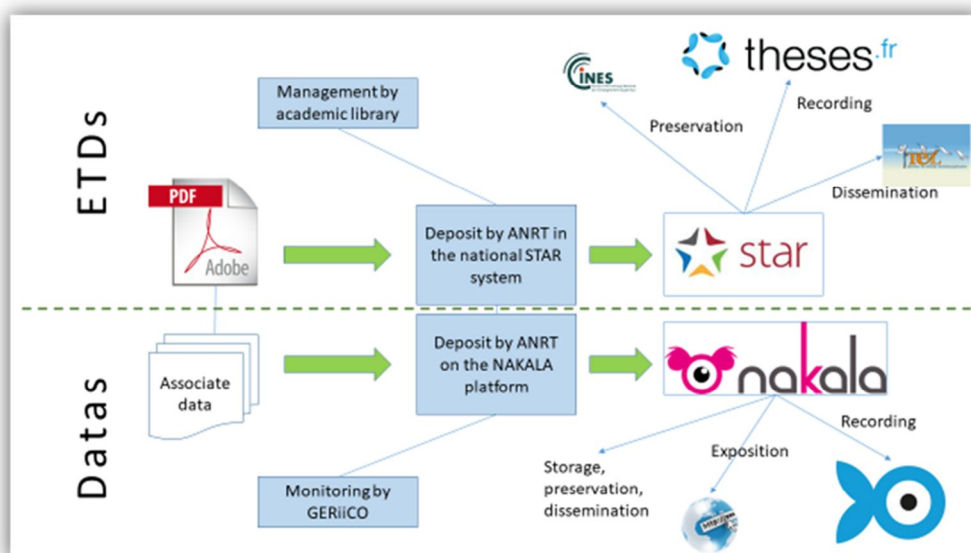


Figure 1: Local ETD/data workflow

Our intention is to offer young scientists a “default solution”, complementary to existing disciplinary data repositories, accompanied by technical assistance and a PhD training program for research data management delivered by our Graduate School.

The preparation and development of the workflow raised several issues, some of them familiar to the grey community:

- Granularity: what exactly should be defined as a dataset for deposit? We have discussed this question in two communications (Schöpfel et al. 2016, 2017a). There are no clear rules or guidelines. The pragmatic solution is to accept datasets on a granularity level which makes sense for understanding (validation) and reuse, and to allow deposit of dataset collections with a hierarchical structure.
- Data structure and description: how are data to be described and structured? Our option is to apply the Metadata Encoding & Transmission Standard of the Library of Congress.

² For instance, the ETDplus project funded by Educopia <https://educopia.org/research/grants/etdplus> and the workflow at the University of Bielefeld, see Vompras & Schirrwagen (2015)

- Identifier: which unique identifier should be used for the datasets? Even if France is part of the DataCite consortium for the assignment of DOIs, we opt for the handle system which is applied by the Huma-Num infrastructure but remain open for future adoption of the DOI.
- Legal aspects: we anticipate legal issues like copyright, third party rights, privacy etc. Our approach is twofold: we provide legal advice as part of the library's data service, and we ask the students to provide a declaration (template) that they have the permission to upload the datasets on NAKALA.
- Quality: the question was raised about the quality of datasets. Should all datasets provided by PhD students be accepted? Should we set up a kind of validation procedure? If so, which criteria should be applied? Who should evaluate? For the moment, we will not filter submitted data files otherwise than by formal criteria (size, format...), similar to other projects and data repositories. But the question remains open.

The tests of the new workflow started end of September 2017. The workflow will be operational in 2018.

Data definition

But what exactly are data and datasets? The issue was raised during the preparation of the data workflow. Therefore, we carry out a conceptual analysis of the meaning and content of the term of research data as a vital complement to the workflow development and survey. The first results were presented during a workshop at the University of Toulouse in May 2017 (Schöpfel et al. 2017a). Figure 2 resumes the main characteristics of our approach which is based on a synthesis of recent French and international reviews and definitions.

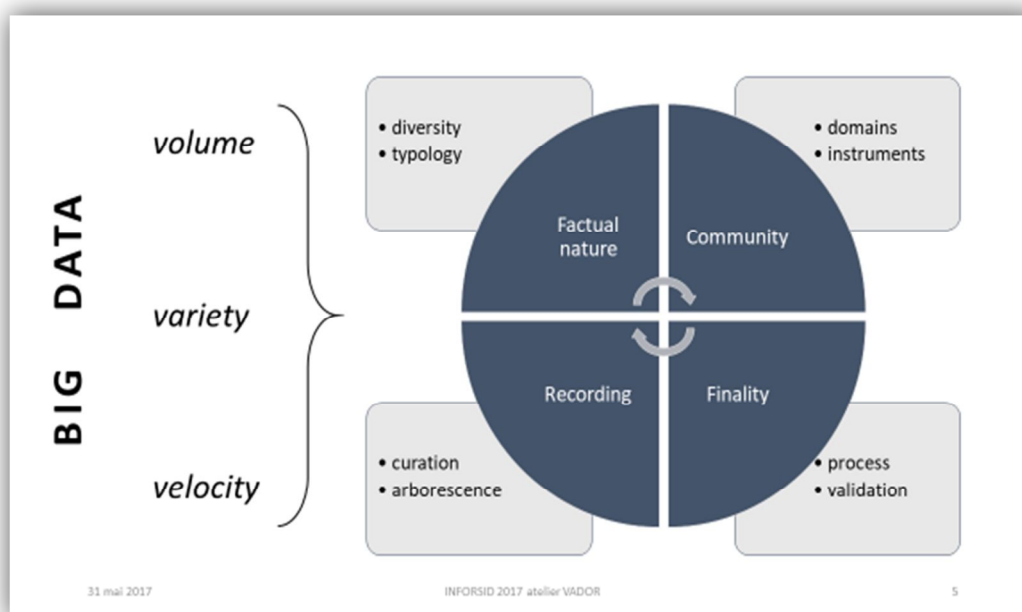


Figure 2: Elements of a data definition

We can identify five key elements of research data:

1. Link to the concept of big data: even if one part of research data is considered as small or “smart” data, the link with the “3 V’s” of the big data is always present³, in particular the diversity of data, their large number and size and the continuous stream of data input and output.
2. Factual nature: definitions of research data often insist on their factual nature, at least implicit, as primary material in need for processing, analysis and interpretation. This often implies a more or less detailed typology of data.
3. The link to community: research data are embedded in disciplinary and institutional context, are specific to large instruments, research infrastructures and scientific domains.

³ See the consensual definition on big data by De Mauro et al. 2016: “Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value”.

4. Finality: research data are also embedded in the research process (cycle), are dynamic, with different functions and requirements. The most basic distinction is between input and output, primary and secondary data, data as resources and data as results of scientific work. Among the various functions, the most important (in a mainly STI and library perspective) are the validation of results and hypotheses (replication) and their preservation along with publications.
5. Recording: the need for recording (curation, preservation) is the last key element of research data definitions. Part of the research data management, data curation raises issues like granularity, identification and data arborescence (hierarchical structure of data and datasets).

Actually, we complete this synthesis with an assessment of the re3data⁴ typology, their distribution and definition especially in the field of social sciences and humanities. Special attention is paid to the content of large data types (raw data, images) and the “other” categories of the more than 500 repositories in SSH.⁵

Data impact: evolution of PhD dissertations

The fourth and last work package of the D4Humanities project is intimately associated with the research and debates in the grey community. Our question is: how does the new environment of research data management and text and data mining impact the characteristics and requirements of ETDs? The discussion is open whether or not PhD dissertations should still be considered as grey literature in the digital age and how (Schöpfel & Rasuli 2017); but it seems obvious that the potential of text and data mining and the availability of datasets related to dissertations will have (or already have) substantial effects on the writing, content, format, length and submission and processing of dissertations, perhaps even on their legal status and licensing.

In the past years, we tried to assess which kind of data are related to PhD dissertations, especially in social sciences and humanities, how they are linked to the dissertation and how they should be curated (Prost et al. 2015, Schöpfel et al. 2015a, b); furthermore, we started to re-examine the meaning of dissertations in the light of text and data mining, considering dissertations as data (Schöpfel et al. 2016). Content mining tends to make the borders between text and data increasingly blurred, even insignificant, and revives the discussion on the distinction between publications (documents) and data. The D4Humanities project contributes to this research field from a special perspective, i.e. the guidelines, prescriptions and laws ruling the writing and submission of digital PhD dissertations. In 2018, the project team will conduct a landscape study together with academic and corporate partners, including a state of the art on recent research and papers on dissertations and data and a small-scale survey on the development of PhD prescriptions.

Perspectives

This last work package is just a beginning. In fact, its objective is threefold:

1. An overview on ongoing research in order to define questions and hypotheses for further research.
2. The setting up of a scientific consortium around a core project team (GERiiCO laboratory at Lille and Institute of Scientific Networking at Oldenburg).
3. And third, the preparation of an international research project on new forms of PhD dissertations, with European (H2020) or French-German funding (ANR/DFG). For the time being, the project’s code name is *xDiss*, for “Special Dissertations”.

Therefore our conclusion is an appeal to the members of the grey community: if you are interested, contact us and join our consortium.

⁴ The international registry of research data repositories, available at <http://www.re3data.org/>

⁵ See Kindling et al. (2017) for some general elements of these repositories. Our own results are stored on a wiki and available on request at <http://d4hdata.pbworks.com>

References

- Chaudiron, S., Maignant, C., Schöpfel, J., Westeel, I., 2015. *Livre blanc sur les données de la recherche dans les thèses de doctorat*. Université de Lille 3, Villeneuve d'Ascq.
- Jacquemin, B., Prost, H., Schöpfel, J., Severo, M., Thiault, F., 2013. Ouvrir les données de la recherche pour la veille scientifique. Le cas des thèses électroniques. In: *VSS'2013*, Nancy, 23-25 octobre 2013.
- Kindling, M., et al., 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* 23 (3/4).
- De Mauro, A., Greco, M., Grimaldi, M., Apr. 2016. A formal definition of big data based on its essential features. *Library Review* 65 (3), 122-135.
- Paillassard, P., Schöpfel, J., Stock, C., 2005. How to get a French doctoral thesis, especially when you aren't French. *Publishing Research Quarterly* 21 (1), 73-93.
- Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. Opening data in PhD dissertations in social sciences and humanities. *Journal of Librarianship and Scholarly Communication* 3 (2), eP1230+.
- Prost, H., Schöpfel, J., 2015. *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3*. Rapport final. Université de Lille 3, Villeneuve d'Ascq.
- Schöpfel, J., 2013. Adding value to electronic theses and dissertations in institutional repositories. *D-Lib Magazine* 19 (3/4).
- Schöpfel, J., Prost, H., 2013. Degrees of secrecy in an open environment. The case of electronic theses and dissertations. *ESSACHESS - Journal for Communication Studies* 6 (2 (12)).
- Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., Thiault, F., 2014. Open access to research data in electronic theses and dissertations: An overview. *Library Hi Tech* 32 (4), 612-627.
- Schöpfel, J., Juznic, P., Prost, H., Malleret, C., Cesarek, A., Koler-Povh, T., 2015a. Dissertations and data (keynote address). In: *GL17 International Conference on Grey Literature*, 1-2 December 2015, Amsterdam.
- Schöpfel, J., Prost, H., Malleret, C., 2015b. Making data in PhD dissertations reusable for research. In: *8th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic.
- Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., Grabbe, P., 2015c. A French-German survey of electronic theses and dissertations: Access and restrictions. *D-Lib Magazine* 21 (3/4).
- Schöpfel, J., Kergosien, E., Chaudiron, S., Jacquemin, B., 2016. Dissertations as data. In: *ETD2016*, Lille 11-13 July 2016.
- Schöpfel, J., Prost, H., 2016. Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas* 29, 98-112.
- Schöpfel, J., Prost, H., Rebouillat, V., 2016. Research data in current research information systems. In: *CRIS 2016*, St Andrews, 8-11 June 2016.
- Schöpfel, J., Kergosien, E., Prost, H., 2017a. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. In: *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017*, 31 mai 2017 Toulouse (France).
- Schöpfel, J., Prost, H., Malleret, C., 2017b. Research and development in the field of research data and dissertations. The D4Humanities project at the University of Lille (France). In: *10th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 19 October 2017, Prague, Czech Republic.
- Schöpfel, J., Rasuli, B., 2017. Are electronic theses and dissertations (still) grey literature in a digital age? a FAIR debate. *The Electronic Library* 35 (4).
- Vompras, J., Schirrwagen, J., 2015. Repository workflow for interlinking research data with grey literature. In: *8th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic.
- Wilkinson, M. D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, sdata201618+.

Public Access to the Dissertations in Russia *

Yuliya B. Balashova <https://orcid.org/0000-0002-2315-8076>
(Russian Federation)

Abstract

To date, a highly constructive situation has developed in the Russian public space in terms of access to dissertations. Mostly thanks to the efforts of official authorities, the procedure for defending dissertations becomes so public for the first time after a long period. At the same time, increasing publicity contributed to the consolidation of civil society and representatives of the scientific community efforts. Accordingly, open access resources containing information about the dissertation thesis, and its author are affiliated both with official and commercial and public organizations.

Keywords: *dissertation thesis, grey literature, science communication, Russian scientific public field.*

Introduction

In the modern Russia as a part of the global world science communication develop quite actively. Its growth is caused, among other things, by the task of wide audience familiarizing with the scientific thought achievements. New popular science media are being created, as well as aggregators of scientific news, scientific festivals are held, science promoted through the social networks.

At the same time, the system of training scientific staff and awarding academic degrees in Russia is still poorly integrated into the world community. The problem is that this system is hybrid, combining Soviet and modern Western properties. From the Soviet times, a complex two-level system of scientific degrees has been existing.

In the USSR, science was respected, partly in the ancient, magical sense, on one hand, from the state, and, on the other hand, from society. Obtaining a scientific degree opened many social possibilities.

The Russian degree of "Candidate of Sciences" corresponds to the generally accepted Ph.D. This degree is considered as the third level of a multilevel system of higher education and is awarded as a result of mastering the doctoral educational program accepted in the West, which in Russia is called "graduate school".

In some European countries awarded the highest scientific degree, named "doctor habilitatus" (Dr. habil. = habituated doctor). It gives the right to occupy a professorship at the university and corresponds to the Russian degree of the "Doctor of Sciences". In order to count on a scientific career in Russia, it's necessary to defend not one, but two dissertations.

In different countries, the state has a different power in regulating the procedure for awarding academic degrees, and accredited doctoral (Ph.D.) programs. In such European, and non-European countries, as France, Norway, Spain, Sweden, United Kingdom, New Zealand, state, and authorized with the state structures play a crucial role in this process. In the USA, and Canada, in many ways existing in a single space, this role belongs to the socio-professional organizations. Finally, in Netherlands the Academy of Sciences has the right to make decision. The right to assign academic degrees is different too. In Russia, and post-Soviet countries, this right is reserved for the state; in the rest of the world – for the universities themselves. Accordingly, in different countries the information support of the dissertation's defense and public access to them are formed in various ways. In Russia, the main role in this process belongs to the state.

Discussion

1. Obligatory stages of the dissertations publicity before defense.

1.1. Publication on the website of the organization, where the work was prepared. There are published the full text of the dissertation, the author's abstract of the dissertation, information about official opponents, and the leading organization, preparing review for the dissertation, as well as the place and time of dissertation's defense. From 2012, at Saint Petersburg State University all dissertations defenses are accompanied by online translation on the university website (Ph.D. SPbSU, 2017).

1.2. Publication on the website of High Attestation Commission under the Ministry of Education, and Science of the Russian Federation. Since 2006, High Attestation Commission publishes on its website the same information, only in a reduced amount, and with reference to the organization's website (High Attestation Commission, 2017)

1.3. The main results of the dissertation research should be published in the form of science papers in the list of peer-reviewed scientific journals. High Attestation Commission on a regular basis revises this list.

2. Obligatory stages of the dissertations publicity after defense

2.1. Obligatory dispatch of the dissertation abstract to the leading libraries of Russia, and also National Library of Belarus (National Library of Belarus, 2017).

2.2. The full written text of the dissertation is sent to the Russian State Library. The main library of the country provides public access to the digitized abstracts, and thesis themselves. Site navigation is available in English (Russian State Library, 2017).

2.3. A mandatory copy of both the thesis and the author's abstract is sent by the regular mail to the Russian Book Chamber (Russian Book Chamber, 2017). In accordance with law, a certain number of copies of any printed publication (books, brochures, periodicals, dissertations, musical notes, geographic maps and atlases) must be sent to the Russian Book Chamber.

3. Other (additional) online channels for dissertations distribution

3.1. Official state resources. The most important is Scientific Electronic Library (eLIBRARY.RU). Materials placed in this electronic library included into the Russian Scientific Citation Index. This Index is a National bibliographic database of scientific citations, accumulating more than 12 million publications of Russian scientists, as well as information on citing these publications from more than 6,000 Russian journals. Other resource, having a printed version, is network encyclopedia "Famous Scientists" (Famous Scientists, 2017). This is a biographical data of scientists, a project of the Russian Academy of Natural History. It also published the main results of the dissertation researches.

3.2. Commercial resources. To such belongs "Library of Dissertations" (DsLib.net). It provides both free and paid access to the thesis text, publishes information about upcoming dissertation defenses. At the same time, monetary deductions to authors of dissertations are declared, but in reality they do not work. We can give another example. "Cyber Leninka" (CyberLeninka, 2017) positions itself as a scientific open access electronic library, the main tasks of which are popularization of science and scientific activity, public control of the quality of scientific publications, the development of interdisciplinary research, the modern institute of scientific review, and the increasing citation of Russian scientists. "Ciber Leninka" is built on the basis of the open science paradigm.

The fundamental problem is that thesis and abstracts both are published "on the rights of the manuscript", in other words, they don't formally have copyright subject. So, private distributors don't pay in fact any royalty that violates the norms of morality, but not the law.

3.3. Public organizations resources. The main source of exposing falsified dissertations is network "Dissernet". It positions itself as "a free network of experts, researchers and reporters, who dedicate their work to exposing scammers, falsifiers and liars. The participants of the community are cooperating with the joint efforts based on the principles of the network distribution of labor and using modern computer technologies, to counteract illegal frauds and infringements in the field of scientific and educational activities, especially in the process of defending dissertations and appropriating academic degrees in Russia" (Dissernet, 2017). More than 10,000 dissertations have been analyzed by "Dissernet", since 2013. Based on "Dissernet" requests, High Attestation Commission made a decision on deprivation of academic degrees. The latest scandal was the case, initiated by "Dissernet", about deprivation of the Minister of Culture of the Russian Federation Vladimir Medinsky academic title of Doctor of Historical Sciences. His dissertation was devoted to the problem of objectivity in the Russian history coverage by foreigners. The main argument of the expert community was that Medinsky has considered the methodology of the historical research as corresponding to the national interests of the state. With this

approach, historical science turns into a servant of the dominant ideology, as it happened in Soviet times. This approach is typical for historical pseudoscience. Nevertheless, at the final meeting on this case, which took place on October 20, 2017, the presidium of the High Attestation Commission retained doctoral degree to Minister V. Medinsky. Such disclosures of the political establishment representatives often occur in Germany (thanks to “VroniPlag Wiki”, as well). And this is an indicator of the healthy society. In Russia, this case additionally indicates that the scientific reputation itself becomes irrelevant, which means a huge problem.

Conclusion

The most positive effect of the current situation is that never before science in Russia has been so public. This result was achieved due to the efforts both of the state (its scientific policy) and society. The next step will be integration this openness into the Western system.

But, in the present time, the procedure for awarding academic degrees begins to change. A number of leading Russian universities and scientific organizations in the near future will receive the right to self-award scientific degrees. On the one hand, thus, Russian universities become more integrated into the Western system. The supervising role of the state will be cancelled; the dissertational councils will not act on a permanent basis, but will be created in accordance with the subject matter of the concrete work. The defenses will take place in English, and – the most important think – the academic degree will be awarded directly. But such innovations cause a lot of criticism.

It is believed that in a democratic society decentralization is extremely constructive. In the scientific policy of modern Russia, this kind of decentralization causes great concern about maintaining a high level of publicity. Actually, there is a return to the closed Soviet system of leading universities.

One of the most serious concerns: will not the thesis be less public, and the thesis defense process more closed?

Acknowledgment

The author gratefully appreciates the support of the Russian Foundation for Basic Research (RFBR).

Funding Information

Article is prepared with financial support of the Russian Foundation for Basic Research. The project No.16-03-50128.

Ethics

This article is original and contains unpublished material. The authors confirm that there are no ethical issues involved.

References

- CyberLeninka. Accessed October 2017. <https://cyberleninka.ru/>
- Dissernet. Accessed October 2017. <https://www.dissertnet.org/about/>
- Famous Scientists. Accessed October 2017. <https://www.famous-scientists.ru/about/>
- High Attestation Commission. Accessed October 2017. <http://vak.ed.gov.ru>.
- Library of Dissertations. Accessed October 2017. <http://www.dslib.net/>
- National Library of Belarus. Accessed October 2017. <http://www.nlb.by/portal/page/portal/index>.
- Ph. D. SPbSU. Accessed October 2017. <https://disser.spbu.ru/ph-d-spbsu.html>.
- Russian Book Chamber. Accessed October 2017. <http://www.bookchamber.ru>.
- Russian State Library. Accessed October 2017. <http://www.rsl.ru>.
- Scientific Electronic Library. Accessed October 2017. <https://elibrary.ru/defaultx.asp>.

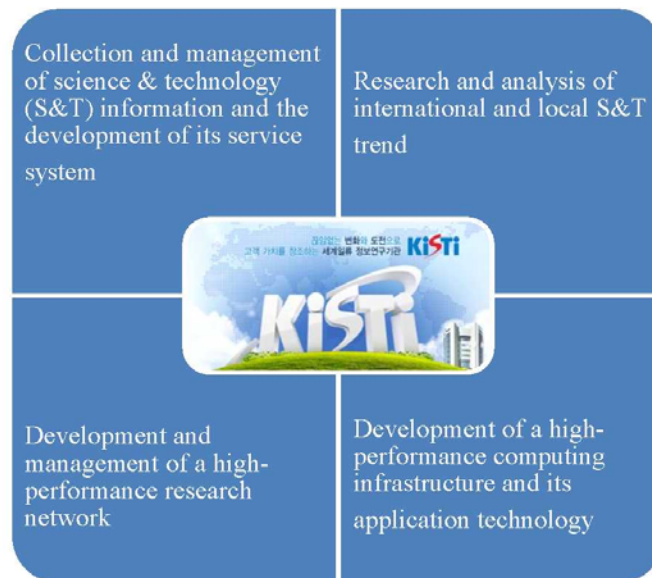
Korea Institute of Science and Technology Information (KISTI)

English version - <http://en.kisti.re.kr/>

* Vision

World-class information research institute creating values for customers

* Main functions



* Management and service of Korean R&D reports

KISTI exclusively manages, preserves, and serves Korean R&D reports for citizens and government officials. It provides Korean R&D reports and their information with National science & Technology Information Service (NTIS) and National Discovery for Science Leaders (NDSL).

* Contact information

KISTI email address: hcpark@kisti.re.kr

Headquarters: Tel : +82-42-869-1004, 1234 Fax: +82-42-869-0969

Data on Grey Literature Containing Scientific And Technical Information*

Leonid P. Pavlov (Russian Federation)

Abstract:

The specificity of data finding and collecting in the field of grey literature (GL) and, in particular, GL devoted to scientific and technical information is characterized. The sources of data concerning the input flow of documents (scientific and technical reports and dissertations) to the Federal GL Information System and covering the years of 1969 – 2018 are described. The System functions in Russia and its databases contain documents in the Russian language. The data is open for a potential reuse once properly obtained from the System.

Keywords: *scientific reports; dissertations; input flow; numerical data*

Subject Area: *Scientific and Technical Information*

Data finding methods

What is generally understood by data papers or data articles is not straightly applicable to the data and datasets used in publications of Scientific and Technical Information (STI) field. There is no "raw" empirical or statistical datasets there typical for many natural and other sciences that are, once obtained, are stored or archived in separate sets for future usage. In case of grey literature (GL) with STI content like scientific reports and dissertations (theses) original data or, rather, information is contained in the databases of abstracted or full text documents stored and archived in digital form. By means of the Federal GL information system on scientific and technical reports and dissertations functioning in Russia and giving a rich set of computer and network services a researcher can find whatever data about this collection he/she needs including content information as a result of a subject search (say, using keywords) or any factual accounts and statistics about the authors, organizations, universities, time periods, financial funding, absolute and relative amounts, regions, etc. Thus, a researcher obtains his/her subject specific dataset from the computer retrieved documents each time he/she gets down to data analysis, description or reviewing.

Collection description

A good example of the scientific paper based on rich numerical data partly given in the tables of the paper is the "Federal GL System Input Flow Analysis" presented at GL15 Conference in Bratislava [1]. The data analyzed in the paper covers the period 1969 – 2012. Selected characteristic years were chosen for the analysis and explanation of the document input flow changes. The complete numerical data for the amounts of arriving documents can be found in different forms depending on the sub-period.

The first sub-period (1968 – 1982) datasets were contained in the "oldest" grey literature databases supported by mainframe computers and stored on magnetic tapes of now exotic hardware, software and data formats. Because of the rapid and cardinal changes in computer technology the databases of this period were not preserved in electronic (digital) form but the documents and some statistics are available in printed form from the then published abstract journals and information bulletins. The qualitative data for this period reflecting the arrival flows of such documents like registration cards, reports information cards, dissertations (theses) information cards, full-text scientific and technical reports, full-text dissertation is published (in printed form, on paper) in [2]. In fact, there is data for the years of 1968 – 1996 printed in this journal. Beginning 1982 and later on, the data of the kind can be obtained from the computer stored databases (in digital form) by means of computerized information system.

This second sub-period (1982 – 2018), in turn, is to be divided into two time spaces: 1982 – 2003 and 2004 - up to now. The information for the 1982 – 2003 period is stored in the databases that contain bibliographic descriptions and abstracts of the documents (reports and dissertations) and is available for online search. The access to the full-text documents of this period is available on the computers of the CITIS reading room in Moscow where a user is able to read them and find facts and figures for his/her research and data analysis. Beginning 2004 and now the full-text documents have been loaded into databases of the modern Federal GL System and available online to authorized users and researches on the Internet site <https://www.rosrid.ru/> (with authors' consent). In case there is no authors' consent

the abstracts of the documents are available for online Internet search and analysis and the full-text documents can be read in the CITIS reading room.

Potential Reuse

Once obtained and published like, say, in [1] or [2], the data and datasets are open for any kind of further research, analysis and reuse with scientific and/or educational purposes. Moreover, a new, more detailed and/or updated data is open to be found in the databases by means of the Federal GL System. The only barrier in this case is that the researcher is supposed to know the Russian language or, as a last resort, Cyrillic letters.

References

1. Starovoitov A.V., Bogdanov Yu. M., Bastrykin A.M., Borzykh A. I. and Pavlov L.P. Federal GL System Input Flow Analysis. Proceedings 15th Internat. Conf. On Grey Literature, 2-3 Dec. 2013. Amsterdam: TextRelease, Dec. 2013. – GL conf. series, No.15.
2. Braguina G.A., Ermakov Yu.S., Pavlov L.P. et al. Technological Processes in VNTICentre. “The Information Bulletin” № 1, July 1997, Moscow, VNTICentre Publishing. Pp. 30-32 (In Russian).

Indexing grey multilingual literature in General Practice: Family Medicine in the Era of Semantic Web^{*}

Marc Jamouille <https://orcid.org/0000-0003-0479-3856> (Belgium)

Miguel Pizzanelli (Uruguay)

Elena Cardillo (Italy)

Ashwin Ittoo and Robert Vander Stichele (Belgium)

Melissa P. Resnick (United States)

Julien Grosjean and Stefan Darmoni (France)

Marc Vanmeerbeek (Belgium)

Problem/Goal: Sharing the results of research with General Practitioners (GPs) is crucial for the survival of the discipline of General Practice / Family Medicine (GP/FM). The production of abstracts in GP/FM probably exceeds 15,000 per year worldwide. Each abstract often represents two years of work for its authors and is expressed in local languages. Only 45% of them are published in indexed medical journals. Usual indexing systems like MeSH are not multilingual nor adapted to the particular field of GP/FM. Consequently, these abstracts are lacking bibliographic control and more than half of the research presented by GPs at congresses is lost. Considering the absence of appropriate domain-specific terminologies or classification systems, we propose a new multilingual indexing system. The existing International Classification of Primary Care (ICPC) is currently used for clinical purposes and has now been expanded with a taxonomy related to contextual aspects (called Q-Codes) such as education, research, practice organization, ethics or policy in GP/FM, currently not captured. The set is proposed under the name Core Content Classification in General Practice (3CGP). The aim is to facilitate indexing of GP/FM specific scientific work and to improve performance in information storage and retrieval for research purposes in this field.

Research Method/Procedure: Using qualitative analysis, a corpus of 1,702 abstracts from six GP/FM-related European congresses was analyzed to identify main themes discussed by GPs (e.g., continuity, accessibility or medical ethics), handled in a domain-specific taxonomy called Q-Codes and translated in 8 languages. In addition, a methodology for building a lightweight ontology (in OWL-2) was applied to Q-Codes, adding object and datatype properties to the hierarchical relations, including mapping to the MeSH thesaurus, Babelnet (www.babelnet.org) and Dbpedia. Finally, the Q-Codes in 8 languages have been integrated in a healthcare terminology service (www.hetop.eu/q) with a companion website (<http://3cgp.docpatient.net>).

Results of the Research: The creation and the on-line publication of this multilingual terminological resource (see <http://3cgp.docpatient.net>), for indexing abstracts and for facilitating Medline searches, could reduce loss of knowledge in the domain. The database is available in 12 human languages and in Web Ontology Language (OWL) for computer use. In addition, through better indexing of the grey literature (online database, congress abstracts, master's and doctoral thesis) of which we show the first results, we hope to enhance the accessibility of research in GP/FM domain and promote the emergence of networks of researchers.

Indication of costs related to the project: This project has not been funded. 3CGP is placed under Attribution-Non-Commercial-Share-Alike 4.0 International (CC BY-NC-SA 4.0). ICPC is copyrighted by WONCA.

Keywords General practice, Terminology, Electronic publishing, Repository, Grey Literature.

1 BACKGROUND

1.1. Need for information in family medicine

In the cycle of patient centered information (Jamouille et al., 2015), the General Practitioner (GP) is simultaneously a heavy user and producer of published/unpublished data. Data could be clinical, (i.e. dealing with symptoms, processes and diseases) or contextual. Contextual data can address particular issues concerning the patient, which may influence the process of care (Schrans et al., 2016). However, contextual data can also deal with issues concerning the doctor, the managerial aspects of care. In this

^{*} First published in the GL19 Conference Proceedings, February 2018.

work, the focus will remain on these last contextual medical features of General Practice / Family Medicine (GP/FM), as its tools for training, research, ethics, inquiry, environmental issues, infrastructure and principle of care.

The realm of GP/FM differs from mainstream health care, as Family Physicians (FPs) address biological, technological, behavioral, sociological and anthropological domains. All of these have a deep impact on the terminologies needed (Helman, 2008; Thompson et al., 2014). The biological and technological fields of medical terminologies are now almost complete (Jonquet et al., 2016; Lelong et al., 2016). However, they sometimes fall short when applied to the field of GP/FM, which relies intensely on complexity and timeline issues (Liang et al., 2014; Madkour, Benhaddou, and Tao, 2016). Albeit well documented clinical issues, professional contextual issues, like management, teaching, research, and ethics are documented in a fragmented way for the first level of care (Jamoulle et al., 2017a).

1.1.1. GP/FM, a profession without clear limits

Despite elaborate definitions of GP/FM Allen et al. (2011) and Primary Care Physicians (PCPs) (AAFP, 2011), the manner in which the profession of GP/FM or PCP is defined and structured varies greatly across family medicine textbooks (Casado Vicente, 2012; Gusso and Lopes, 2012; Kochen, 2012; Murtagh, 2011; Druais et al., 2009; David et al., 2013; Lakhani, 2003; McWhinney, 1997). If one examines the table of contents of these cited works, as far as the general management and the contextual background are concerned, the quoted books are absolutely different. None give a similar view of the scope and contextual scope of family medicine. The GP/FM vocational training programs have no homogeneity, despite the recommendations of EURACT, the WONCA Europe working group on education (Heyrman, 2005). Also, when considering Continuous Medical Education, the drug industry's influence on the choice of subjects is decisive, which creates multiple conflicts of interests (Davis, 2004).

1.1.2. Published and unpublished in GP/FM, what's the meaning?

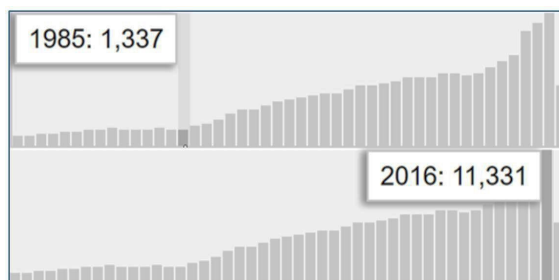
Medical Subject Headings (MeSH) are normalized keywords directed to enter queries in Medline, the bibliographic data base of the National Library of Medicine (NLM) through the interface PubMed (Lowe and Barnett, 1994). Currently the PubMed interface gives access to 27,3 millions of citations. A search with 7 GP/FM and 8 specific MeSH descriptors of Primary Health Care (PHC) gives 480.000 citations. (PHC and GP/FM related MeSH are listed in Fig.1), less than 2% of Medline content.

Figure 1. GP/FM & PHC related MeSH in PubMed

Types	MeSH	Year of introduction
Primary Health care (PHC) related MeSH	"Community Health Service" [MeSH:Noexp]	1967
	"Community Mental Health Services" [MeSH:NoExp]	1967
	"Home Care Services" [MeSH:NoExp]	1967
	"Primary Health Care" [MeSH:NoExp]	1974
	"Community Health Centers" [MeSH:NoExp]	1979
	"Community Mental Health Centers" [MeSH:NoExp]	1979
	"Home Care Agencies" [MeSH]	1995
	"Rural Health Services" [MeSH:NoExp]	1996
General Practice / Family Medicine (GP/FM) MeSH	"Physician, Family" [MeSH]	1974
	"Community Medicine" [MeSH]	1977
	"Family Practice" [MeSH]	1978-2010
	"Gatekeeping" [MeSH]	2000
	"General Practice" [MeSH]	2011
	"General Practitioners" [MeSH]	2011
	"Physician, Primary care" [MeSH]	2011

The rise in publications in GP/FM has been inevitable since 1985 (see Fig. 2). This raises question about the Pisa definition of grey literature, considered as *not controlled by commercial publishing* (GreyNet, 2014), as one can hardly pretend that papers published by the NLM are always commercially controlled.

Figure 2. Evolution of publications in GP/FM (7 MeSH) from 1985 to 2016 on PubMed



Unpublished data in the GP/FM field are numerous. Yet, GP/FM organizations are heavy producers of continuous medical education (VanNieuwenborg et al., 2016). They contribute greatly to training sessions and organize local, regional and national level medical conferences, as well as to research meetings (Buono et al., 2013), virtual conferences (Cavadas, Villanueva, and Gervas, 2010), websites, and blogs. They are also active on social networking (Veuillotte et al., 2015). With so many heavy producers of various nationalities, it is important to note that at local and national events, the local language is the rule.

It is not a secret that knowledge translation between doctor and patient is highly controlled by pharmaceutical companies (Moynihan, 2003) (Moynihan and Bero, 2017). Despite the activities of GP organizations, in some countries, domestic papers as medical newspapers remain the main sources of information for practicing doctors (Tabatabaei-Malazy, Nedjat, and Majdzadeh, 2012). Usually, they are edited within an atmosphere of heavy, silent corruption (Angell, 2017). The translation of information by drug representatives is also a determining factor (Greenway and Ross, 2017) as well as predatory open access publications (Shen and Björk, 2015) or pay for publishing process (Quan, Chen, and Shu, 2017) which dismisses whole sectors of publication. This implies that knowledge management tools in GP/FM must be controlled by dedicated, unbiased GPs. The movement of free lunch doctors (www.nofreelunch.org/ (USA)), the no-gracias one (www.nogracias.eu/ (Spain)), the Medico Sin Marca (Chile), (www.medicossinmarca.cl), the Therapeutic initiative (Canada) (www.ti.ubc.ca) or, more generally, the members of the International Society of Drug Bulletins (ISDD) (www.isdbweb.org) refuse to adhere the pharmaceutical industry influence. However, sharing the results of research with General Practitioners is crucial for the survival of the discipline of GP/FM, which means a universal system must be created (McIntyre et al., 2016).

1.2. Producing Information at the point of care

It could be believed that a simple terminological subset may be sufficient to meet the needs of GP/FM computer systems. Lack of visibility of the complexity of the work of family doctors has allowed for such biased vision. Of course, the family physician often sees simple problems. But he is accompanying a set of patients throughout a lifetime. The family physician knows a patient more extensively than most specialists do. He also becomes a specialist in patients bearing rare diseases and of various cultural backgrounds. He will, therefore, have extensive terminological needs, even more extensive than many specialists.

1.2.1. Clinical information

The adjective clinical deals here with patient related data, such as: reasons for encounter, symptoms, acts performed or requested and diagnosis. Terminology for clinical information is a highly specialized and difficult field of current medicinal research (Jamouille et al., 2014). Clinical information is accumulated in Electronic Medical Records (EMRs) and, if well organized, transferred to study centers where huge database may be used to teach medicine, analyze epidemiological data or be used for secondary searches (Charlton et al., 2010; Carey et al., 2004; Britt et al., 2003).

1.2.2. Professional contextual information

The term *contextual* applies as a generic term for the name of the taxonomic product presented here, by-product of a PhD thesis in medical sciences (Jamouille, 2017). A concept like *uncertainty*, the usual companion of the doctor or the concept of *quality assurance* or *environmental health*, are all essential elements of professional practice. These are not clinical terms as they do not always deal with current patient problems. The term *contextual* appeared the most relevant, as it was defined in the Meriam-Webster Dictionary as: *the interrelated conditions in which something exists or occurs*.

1.3. Consuming information at the point of care

As stated by James (2016), the Internet has triggered a transformational change in the dissemination of science in the form of a global transition to open access (OA) publishing. GPs, the rank and file (Chinitz and Rodwin, 2014) workforce in medicine are using those resources extensively despite sometimes huge material difficulties to access the sources when working in rural or remote areas (Salman Bin Naeem, Shamshad, and Amjid, 2013).

1.3.1. Sources of information at the point of care in GP/FM

On-line, directly accessible major documentary databases in open access and local language are not numerous nor specific to the GP/FM field (Hubbard, 2008). The US NLM PubMed database (<https://www.ncbi.nlm.nih.gov/pubmed/>) gives access to millions of paid and open access publications, mostly in English. The Pan American Health Organization sponsors Scientific Electronic Library on-line (www.scielo.org). SciELO is the South American champion of free access journals, mostly in Spanish (PAHO Bireme Sao Paulo, 2016). The World Health Organization (WHO) supports the Global Index Medicus (www.globalhealthlibrary.net). In France, LiSSa (www.lissa.fr) gives access to more than 10⁶ citations of French literature (Cabot et al., 2017a).

To say nothing of Google Scholar, the web is a considerable resource of information in medicine, especially in gray literature. Janamian et al. (2016) have identified and searched for 260 web sites as GP/FM sources. As stated by González-González et al. (2007) *"In primary care, each practitioner encounters more than 500 different clinical topics in any year"*

1.3.2. Use of Medical Subject Headings (MeSH) descriptors in GP/FM

General Practice is a profession generally regarded as part of the first line of care, PHC, a form of care organization. General Practice and Primary Health Care concepts share the same extension, but not the same intension. The first describes the duty of a profession, the second the management of a service (Jamouille et al., 2017b).

1.3.3. Use of Health descriptors (DeCS)

The controlled vocabulary of Health Sciences Descriptors (DeCS), initially a translation of MeSH into Spanish and Portuguese, has been expanded with new categories. It was also adopted into the indexing and multilingual search of the scientific and technical literature in South America (<http://decs.bvs.br/l/homepagei.htm>).

1.3.4. ICPC for Classifying Clinical Issues in GP/FM

The International Classification of Primary Care (ICPC) (Soler, Jamouille, and Schattner, 2015) is routinely used by physicians around the world to categorize the problems encountered in their practice with patients, i.e. their clinical activity. We will see that ICPC has proved effective in many other uses and show that it is adapted for collecting clinical problems that doctors discuss at congresses.

1.4. Grey Literature as source of information

The price of access to international high-level journals, mostly exclusively in English, is prohibitive. The economic aspect is therefore a major obstacle. However, the change of economic model is under way. McKenzie (2017) considers that the explosion of the use of Sci-Hub facilities (<http://sci-hub.tw>) is the beginning of the end of the scholarly publishing. Whatever the case, the open access grey literature in medicine is in full development (Swan, 2012). According to Schöpfel (2015), *The term gray literature remains ill-defined, imprecise, with fuzzy outlines*. We propose to consider grey literature in GP/FM publications that share the following characteristics:

For the background:

- Sharing knowledge specific to the field of GP/FM, no matter the format (papers, articles, memo, master thesis, PhD thesis, leaflet, abstracts of presentation, web pages, video, images, YouTube, Facebook, Twitter, Google+, LinkedIn, dataset).
- Being unreferenced in well-known local or international medical databases (PubMed, LiSSa, Scielo, Lilacs, ORBI, etc.).
- Being submitted to a scientific quality assurance process (in anthropology or bio-sciences).

For the format:

- Being freely accessible in an Open access model.
- Using a systematic multilingual vocabulary encoding scheme (indexing system).
- Relying to Dublin Core Metadata Initiative or equivalent standardization process.
- Being ready for machine use in the semantic web.

1.4.1. Metadata and Vocabulary Coding Scheme

Metadata consists of statements we make about resources to help us find, identify, use, manage, evaluate, and preserve them (Sutton, 2007). Metadata may be interpreted by machines and people. Dublin Core Metadata Initiative (DCMI) (<http://dublincore.org/>) provides simple standards to facilitate the finding, sharing and management of information. Metadata are basic description mechanism for digital information that, can be used in all domains, for any type of resource, simple, yet powerful, can be extended and can work with specific solutions, making it easier to find information on the Web as it develops.

1.4.2. Grey literature and Semantic web opportunities

Metadata allow the retrieval from data from dedicated repositories. Nevertheless, as stated by Goggi et al. (2015), *documents may contain important information that has not been encoded in the metadata*. Extracting key concepts from unstructured texts is the following step, done by semantic annotators, by-product of research in Natural Language Processing. Key concepts could be added to indexing facilities or tagged as identifiable information for use in Linked Open Data (LOD). This opens the *possibility of enhancing the visibility and accessibility of grey literature via its connection to the data it describes and to an advanced full text indexing* (Goggi et al., 2015).

2 AIM OF OUR RESEARCH: PROPOSAL FOR A NEW CODING SCHEME IN GP/FM

Our work aims to identify the themes in knowledge production by GPs in a new Vocabulary Coding Scheme called Core Classification of General Practice Family Medicine (3CGP). This program encompasses clinical and contextual situations in the GP/FM practice. Simultaneously, we hope to develop our system in such a way that machines (i.e., computer), could deal with that data and reason about it using the Semantic web technologies.

The absence of adapted concepts and descriptors for contextual aspects of GP/FM is one of the reasons why the scientific work of family physicians is hard to retrieve from mainstream bibliographic systems. In addition, more than 50% of the scientific output of GPs at conferences is never published (Van Royen et al., 2010). There are no dedicated indexes of grey literature (Mahood, Eerd, and Irvin, 2014), and abstracts or collections of dissertation titles are often not properly indexed in this field (Lawrence et al., 2014).

This work presents a new taxonomy of contextual aspects of GM/FM, in hopes of helping to improve the situation surrounding GM/FM grey literature. Taxonomies provide schemes to help classify entities and define the relationships between them (Dixon, Zafar, and McGowan, 2007). The purpose of this development is also to provide tools to exploit modern technology - in terms of terminology for information storage and retrieval systems (Vanopstal et al., 2011), such as: machine learning, semantic web techniques, natural language processing (NLP) and linking data. This kind of system is already in use in clinical settings for patient data (Colliers et al., 2016) and one hopes to apply such techniques to an indexing system in a near future for the communication of family doctors in congresses and related grey literature.

In brief, our aims are triple:

- To improve annotation of grey literature in primary care.
- To facilitate indexing of congress abstracts and theses.
- To improve the searchability of repositories for these information artefacts.

3 METHODS

3.1.1. Referring to METHONTOLOGY steps for the development of the project

The phases of development of the project are shown on Fig. 3 along the time line. Qualitative analysis of communications of GPs during congresses has induced the creation of a controlled vocabulary organized in a taxonomy. To develop a domain-oriented taxonomy (the simplest form of an ontology- i.e., a light-weight ontology), methodology for ontology construction was included (Gómez-Pérez, Fernández-López, and Corcho, 2003). The four main phases of the METHONTOLOGY process are shown Vertically:

1. Knowledge Acquisition and formalization;
2. Integration process;
3. Implementation;
4. Publication and Dissemination

Knowledge acquisition, formalization and integration were added in 2005. The implementation phase in the online Hetop server began in 2014. We have added a dissemination phase through Internet and publications (see <http://3cgp.docpatient.net/>)

3.1.2. Knowledge acquisition & formalization; Qualitative analysis of GPs' communications by a Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

Using qualitative analysis, a corpus of 1,702 abstracts from six GP/FM-related European congresses was analyzed by qualitative methods to identify 182 themes discussed by GPs (e.g., continuity, accessibility or medical ethics), handled in a domain-specific taxonomy called Q-Codes and translated into 8 languages. To identify key concepts in a domain-specific taxonomy, data is analyzed in a grounded theory approach (Glaser and Strauss, 1999). The qualitative analysis software (ATLAS.ti® <http://atlasti.com/>) was used. ATLAS.ti enabled the ability to map specific words to already-defined ICPC-2 and to find new concepts to feed the new Q-Codes taxonomy. The same theme could not reappear in the same abstract more than once, and (generally) no more than six themes were identified in each abstract. The analysis performed by EGPRN in 2010 on 614 abstracts (Kruschinski and all, 2010), using a similar approach, has been used to control the QR (Research) domain and check the consistency of the Q-Codes proposal.

3.1.3. Integration phase, birth of 3CGP

The Core Content Classification in General Practice/Family medicine (3CGP) is formed by the addition of ICPC-2 for clinical issues and Q-Codes for professional contextual issues, both discussed during meetings between GPs. The Q-Codes taxonomy was elaborated on the model of ICPC, using the letter Q to categorize the contextual elements, for the letter Q was unemployed in ICPC-2.

$$\text{ICPC-2} + \text{Q-Codes} = \text{3CGP}$$

Q-Codes are born from the qualitative analysis of 1,702 abstracts of conferences. Naturally, we hope that future conferences will allow for a surge in new concepts and new entries in the Q-Codes classification. Q-Codes are divided into 8 domains. The taxonomy starts with the QC domain, which represents *Patient's category*, and covers topics such as age, gender issues, and victim-hood. The second one is the QD domain, representing *Family doctor's issue*, which covers issues such as disease management, communication, clinical prevention, and medico legal issues. QE represents *Medical Ethics*. This domain covers bioethics, professional ethics, and info-ethics. The fourth domain is QH, representing "Planetary Health", which deals with such areas as environmental health, biological hazards, and nuclear hazards. The fifth domain is QP, *Patient Issue*, which includes patient safety, patient centeredness, and quality of care. The QR domain is *Research & Development*, covering research methods, research tools, and epidemiology of primary care. QS is the Structure of Practice domain. It covers topics such as primary care settings, primary care providers, and practice relationships. Finally, the QT domain is Knowledge management. This domain deals with teaching, training, and knowledge dissemination. Each domain of Q-Codes is divided in Categories, Sub-Categories, and Sub-Sub-Categories. A ninth domain, QO for Other will be used in the abstract coding process for not precise descriptions or for a concept worth to be considered as a potential candidate for a new theme.

Figure 3. The phases of development of the project on a time-line. The four main phases of the METHONTOLOGY process

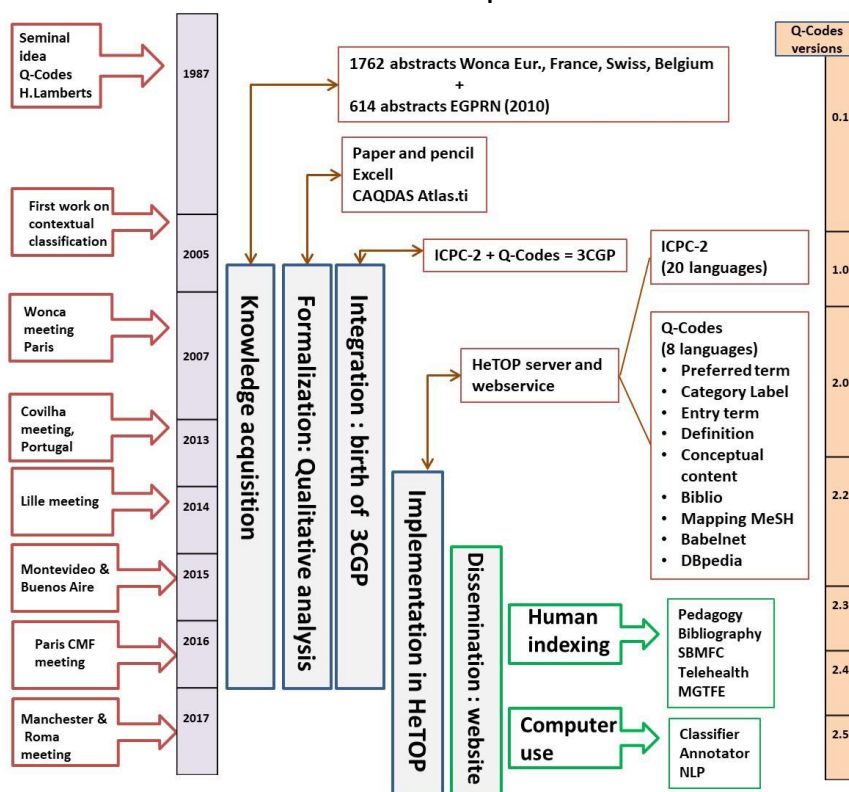


Figure 4. Congress CNGE 2013: After being scanned, coded themes appear on the right column: Q-Codes (QS41, QC3, QP31) and one ICPC code (Z01). Here, the theme identified is the offer of family medicinal services in vulnerable populations. At the bottom left, the software proposes the list of pre-registered Q-Codes. (ATLAS.ti Software)

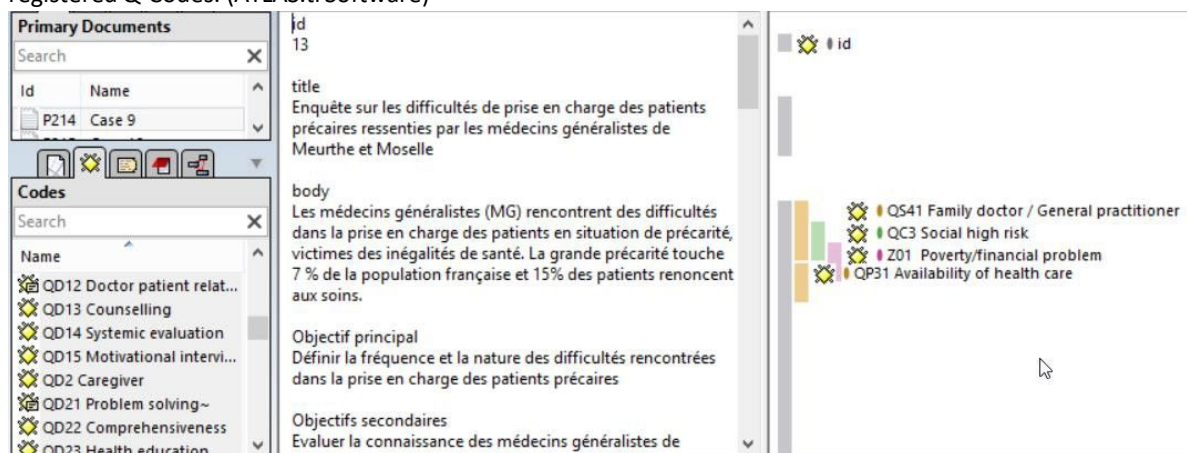
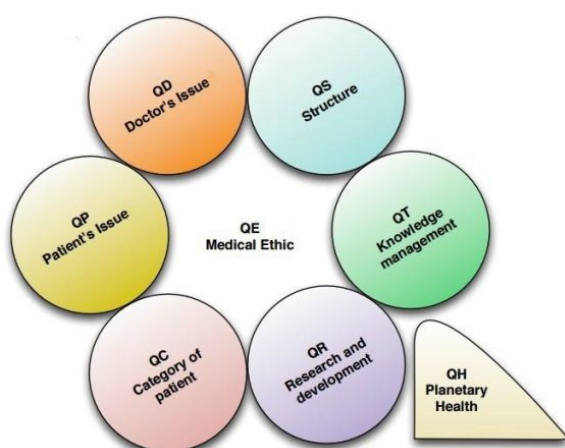


Figure 5 The Q-Codes matrix in the shape of a Q-Letter

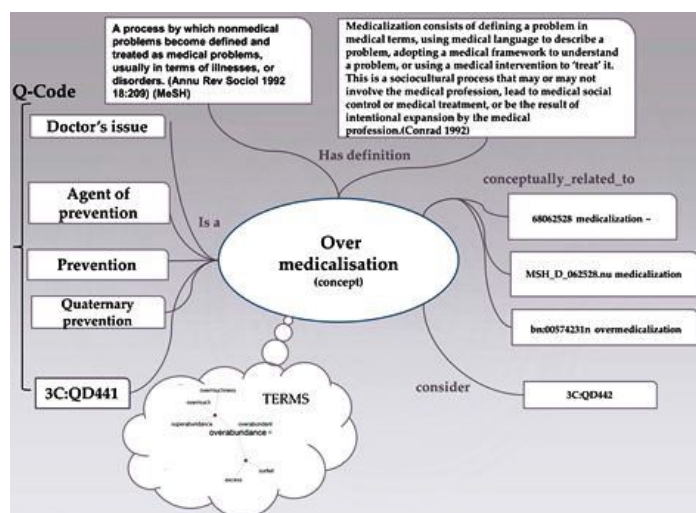


The presentation of the Q-codes under a matrix format is shown on Fig. 5. The matrix takes the shape of the letter Q, representing the 8 domains of the Q-Codes. On the left, the people related domains - Doctor's issue, Patient's issue and Category of patients; On the right, the managerial related domains - Structure, Knowledge management including Teaching and Training, and Research and development; Hazards are the underlying Planetary health conditions represented by the downward oblique tail stylized as a triangle but which are in reality the back-ground of the GPs work; in the center, joining all, Medical Ethics. Note that the Q's tail, which is the Planetary Health, prevents the wheel from turning endlessly. This is a nice demonstration of the importance of the environment on health issues (Graphic design Patrick Ouvrard).

3.1.4. Implementation ; Organizing the concepts of the taxonomy following a Data Structure Diagram on the HeTOP server

Integration and implementation came to fruition in the meantime. The ICP-2 classification was edited on the HeTOP web site in 22 languages and the Q-Codes in 12 languages; French, English, Dutch, Spanish, Portuguese, Vietnamese, Turkish, Georgian, Greek, Korean, Italian, Ukrainian. The Data Structure Diagram, a graphic technique, based on a type of notation dealing with classes of entities and the classes of their relationships (Bachman, 1969) has been used to organize the mappings. In Fig.6, the central concept (herē, Overmedicalisation), is linked by its relations (is a - consider - has a definition, conceptually related to) to other formally defined fields of knowledge. This kind of structure is machine readable and forms the basic structure of our taxonomy.

Figure 6. Data structure diagram (DSD) of a Q-Code, showing the map of concepts and their relationships (conceptual data model)



3.1.5. Dissemination; The HeTOP server as a GP/FM knowledge resource

The HeTOP server, produced by the Department of Medical Information and Informatics (D2IM) of Rouen University (France) is edited in the Web Ontology Language (OWL) (McGuinness and Harmelen, 2004). HeTOP is based on a multi-terminology meta-model and is cross-lingual since terminologies and ontologies are often available in several languages. The web site can be used by both humans and machines via a dedicated web service (www.hetop.eu). Each HeTOP rubric could be also expressed under an Unique Ressource Identifier (URI) format. Adding Q-Codes to ICPC-2, it was then possible to develop a complete terminology adapted to GP/FM and PHC needs. Although WONCA retains copyright on the use of ICPC, every effort should be made to disseminate it. The HeTOP database is freely accessible by means of a simple registration process. The Q-Codes belong to the author, but they are licensed under a Creative Commons Attribution Non Commercial (CC-BY-NC) licence.

3.1.6. Websites

Free access, bibliography & user guide can be found at <http://3cgp.docpatient.net> . On this site one can reach the 3CGP HeTOP pages i.e. the Q-Codes & ICPC-2 repository at Rouen University Hospital, France in English – Español – Ελληνικά (Greek)– Français – ქართული (Kartouli) – Italiano – 한국인 (Korean)– Nederlands – Português – Türk – Українська (Ukrainian)

4. RESULTS

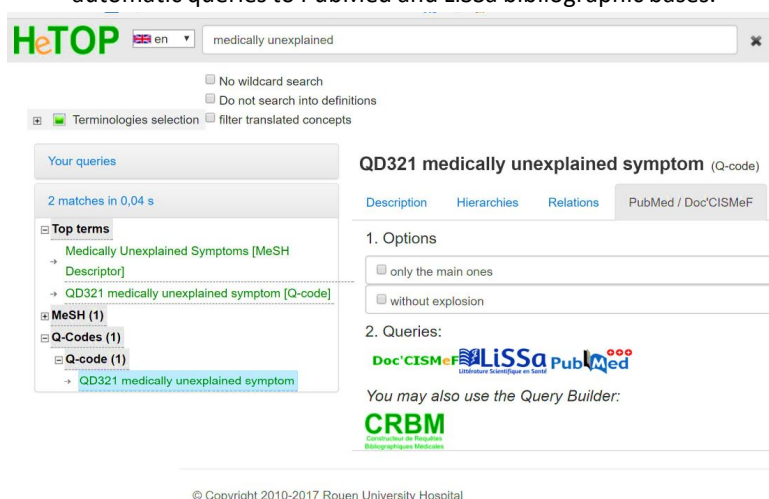
4.1.1. Pedagogical use

The ICPC is used worldwide as the main data producing system in Primary Care. It is incorporated into Health Information Systems and used in Electronic Medical Records in numerous countries. Availability of ICPC-2 in multilingual URIs is a must for teaching ICPC worldwide. The eight domains addressed by the Q-Codes are the embryo of what could become the table of contents of GP/FM. Teaching GP/FM is a must when referring to a rarely taught although so frequent as *Medically unexplained symptoms* or *Indoor pollution*. The terms and definitions of the 182 Q-Codes are available in multiple languages, stressing the international interest surrounding this database. The terms and definitions have been edited in book format in 6 languages (es, pt, fr, en, nl, vi). All versions minus Vietnamese are available at the printing office (<https://www.publier-un-livre.com/en/>).

4.1.2. Bibliographic use

GP/FM has no specific indexing system. Twenty per cent of the ICPC-2 codes and all the Q-Codes are mapped automatically to MeSH and each mapping curated manually to MeSH of the National Library of Medicine. Q-Codes are a wonderful tool for teaching specific fields of GP/FM. They are also a useful resource of knowledge for students, researchers and working practitioners at the point of care. Automatic specific citations retrieval system allows access to dedicated bibliography on PubMed but also to LiSSa, the French resources base in medicine (see Fig. 7).

Figure 7. The HeTOP query interface for the Q-Codes *Medically Unexplained Symptom* proposes automatic queries to PubMed and LiSSa bibliographic bases.



© Copyright 2010-2017 Rouen University Hospital

The power of Q-Codes as bibliographic tool is also used in a Collaborative Database on Quaternary Prevention Resources and References carefully described online in four languages (<https://tinyurl.com/Tutorial-P4-Library>) The current platform has been designed to be augmented from various sources and shared with whomever necessary. The documents and materials that are uploaded to the platform can be shared through an Excel document. The entries have been coded with the multilingual terminology of general practice and family medicine, Q-Codes (<http://3cgp.docpatient.net>)

4.1.3. Indexing master theses

Figure obtained by manual indexing of conferences to develop the Q-Codes taxonomy is still a source of information on the interest of participating GPs. In Fig.8 we show the distribution of the rubrics of the QT domain in a French congress of GPs. In this domain, several experiments are ongoing. The work of doctors in general and family medicine training is often of high quality, requires considerable investment from the authors and sometimes represents little-explored research areas. The leaders of the association of the three Belgian francophone universities (CCFFMG) (see in French; www.mgtfe.be/le-guide-dindexation-dun-tfe/) decided to give visibility to this work by publishing the best of them online. The 3CGP indexing system was chosen to index work by authors At the time of registration. This guide is also available online in English (<https://tinyurl.com/Q-Codes-guide>). The figure 9 shows the results of the coding of student thesis with Q-Codes at Coimbra university medical school between 2008 and 2017. Patient issues (QP) are of growing interest while Research, environment and Ethics are nearly absent. (de Oliveira Tavares & all 2018)

Figure 8. Distribution of QT (Training & Teaching) codes in a congress of French GPs showing the main domains of Interest of participating doctors (number of codes used on 212 abstracts) (Q-Codes version 2.3).(Data Marc Jamoulle)

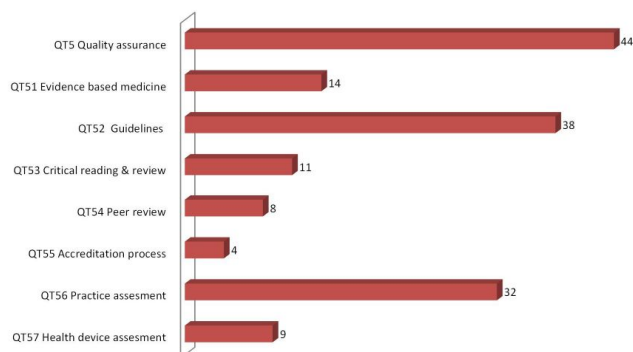
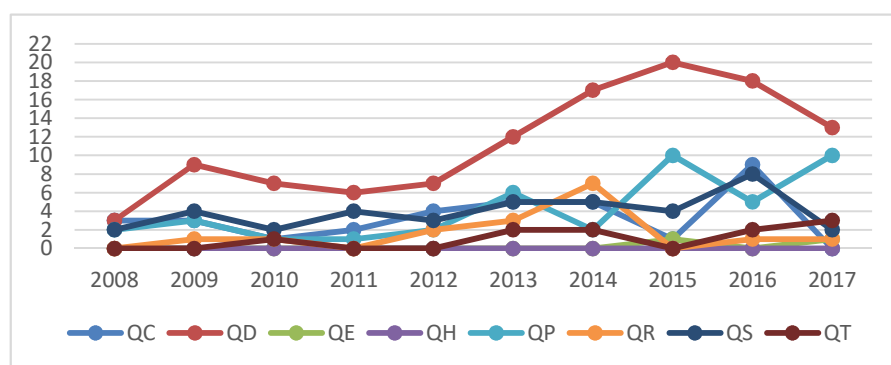


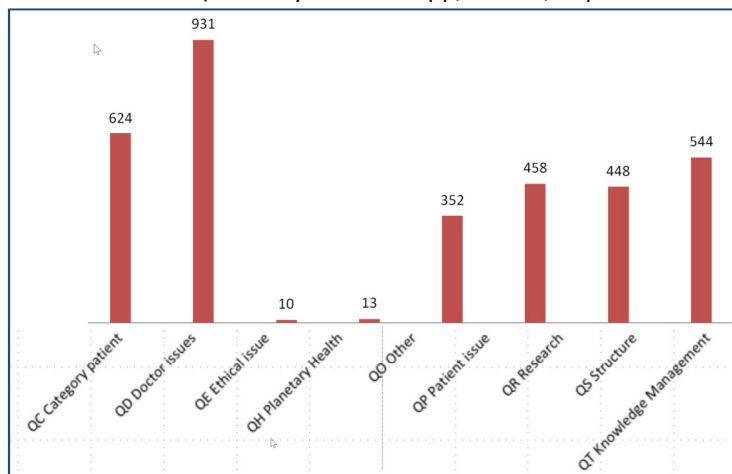
Figure 9 Evolution over time of the contextual contents of the 3CGP in the 6th medical year final assignments. QC-person's category; QD-doctor's issue; QE-medical ethics; QH-planetary health; QP-patient issue; QR-research; QS-structure of practice; QT-knowledge management (Courtesy Dr de OliveiraTavares)



4.1.4. Indexing of congresses

The system developed to index master's theses is reused by the Brazilian Society of Family and Community Medicine (SBMFC). The 3CGP coding system is in use at the deposit page of their 14th Congress. The participants have to choose at least two codes and a maximum of 4 codes of ICPC-2 and Q-Codes. 1,746 reviewed and coded abstracts by participants with ICPC and Q-Code have been retrieved and analyzed. The same experience is planned in 2019 during the next congress.

Figure 10 Distribution of the coding by the participants of abstract presented during the Curitiba meeting, Brazil on Family Medicine in 2017. On a total of 1746 communications, only 10 was dealing with ethics and 13 with environmental issues while 931 was about Doctor's issues or 448 about Structure of care. (Courtesy Daniel Knupp, SBMFC, unpublished data)



4.1.5. Indexing question-answer pairs

This approach has been used on data from Pernambuco, Brazil, for initially manually indexing a sample of 550 questions; with an ultimate goal of semi-automated indexing of larger data sets, measured in the tens of thousands of question-answer pairs. These question-answer pairs originate from the Brazilian Telehealth system, representing communication between rural health care providers and nurses and doctors in the urban Telehealth centers. (Resnick et al. 2013)

4.2 3CGP use by machines

4.2.1 Automated classifier

To find an automated method capable of analyzing the content of non-clinical General Practice articles and predicting the corresponding Q-Codes categories is not an easy task. A classifier was already developed at the department of Information Systems at the University of Liège (HEC, Professor Ashwin Ittoo). The main difficulties arose from the small amount of sample data available, the large number of categories to be identified, and the high specificity of the scope of Q-Codes making categories difficult to discern (Rigaux, 2015). The classifiers also use filtered lemmatization, and they obtain a modest F1-score of 0.452 and 0.344 respectively. The full work is available in French on; <https://tinyurl.com/yc5ej2bw>.

4.2.2 Automated annotator

The tool *Extracting Concepts with Multiple Terminologies* (ECMT) is a web service developed at D2IM, Rouen. It aims to fully, automatically identify clinically relevant entities in medical texts in French with several types of documents: abstracts titles, documents about marketed drugs and death certificates. The extraction is performed at the phrase level of the text. ECMT has also a user-friendly interface accessible after authentication (<http://ecmt.chu-rouen.fr/>).

Figure 11. Automatic annotation of concepts by ECMT v3 by MeSH (MSH), National Cancer Institute (NCI), MedDRA (MDR), SNOMED (SNO) etc. The red arrow shows the automated identification of concepts in Q-Codes (CGP); QD4 Prevention and QD44 Quaternary prevention (in French).

Extracteur de Concepts Multi-Terminologique (ECMT v3)			
How-to - Contact - © 2017 CHU de Rouen - CISMef.			
La prévention Quaternaire (P4) est l'ensemble des activités de santé qui atténuent ou empêchent les conséquences des interventions inutiles ou excessives du système de santé.			
Effacer	1 phrases annotées en 525 ms. 44 codes distincts identifiés.		
Terme	Ter. Code		
154.24 - Activités	DEW 154.24	Prévention	TSP 009460
320.101 1 - Systèmes	DEW 320.101 1	Prévention	MDR 10036654
540.113 - Systèmes	DEW 540.113	Prévention	NCI C15843
551.79 - Quaternaire	DEW 551.79	prévention et contrôle	MSH Q000517
Activité	TSP 000206	Prévention santé	TSP 009471
activité	IUP A00113	procédure	SCT 71388002
activité	NCI C43431	procédure préventive (procédure)	SCT 169443000
activité	SCT 257733005	QD4 prévention clinique	CGP QD4
Analyse systémique	TSP 000743	QD44 prévention quaternaire	CGP QD44
Conséquence	TSP 002943	S70BJ01 PREVENTION	CLA S70BJ01
conséquence	NCI C74555	S72EA PREVENTION	CLA S72EA
Ensembl	NCI C45763	Santé	MSH M0009825
ensemble	NCI C63802	Santé	ICN 10008711
ensemble	NCI C47894	Santé	SCT 263775005
excessif	NCI C73992	santé	MSH D006262
excessif	SCT 260378005	santé	NCI C25178
Intervention	RAD RID10381	Système	NCI C25700
Intervention	NCI C25218	système	SCT 246333005
LE systémique	MDR 10024067	système	NCI C40568
médecine préventive	CIS MT21	système	IUP S06234
médecine préventive	MSH D011315	système	SCT 31099001
		système	SNO G-A572
		système	NCI C13310

The pertinent terms are retained based on the HeTOP resources. In Fig. 11 an example is given for the processing of the phrase: *La prévention Quaternaire (P4) est l'ensemble des activités de santé qui atténuent ou empêchent les conséquences des interventions inutiles ou excessives du système de santé.* [Quaternary Prevention (P4) is the set of health activities that mitigate or prevent the consequences of unnecessary or excessive health system interventions.] ECMT extracts the terms in French from terminologies in HeTOP like MeSH or the National Cancer Institute terminology (NCI) as well as from Q-Codes terminology (CGP). As observable in Fig.11, QD4 Clinical Prevention and QD44 Quaternary Prevention have been identified.

4.2.3. Using Q-Codes in an e-learning program, Vietnam

Dr. Thành Liêm Võ, from the family medicine unit of the Pham Ngoc Thach Medical University, Ho Chi Minh City, Vietnam (www.pnt.edu.vn/vi/) has incorporated the Q-codes in Vietnamese in an e-learning system for medical students. The glossary of Family Medicine terminology helps one to understand and standardize the complex concepts of the discipline. It reduces the variety of these interpretations. Vietnamese versions of Q codes are used as a source of reference. For now, Q-Codes has been integrated into the format of glossary in 3 months FM training at Pham Ngoc Thach Medical University.

5 DISCUSSION

5.1 Main findings

Three areas of knowledge are at stake in this study: (i) Family Medicine as a pillar of primary care, (ii) Computational linguistics, and (iii) Information systems. The association of ICPC, in its three components Symptoms, Procedures and Diagnostics, with the Q-Codes forms an indexing system. This system therefore covers clinical and contextual elements specific to General Practice and Family Medicine. This system allows us to identify patients' symptoms and complaints, diagnosis or disease hypotheses, processes used by physicians, either by themselves or by third parties, and, finally, the context of application given by Q-Codes.

The Q-Codes represent a form of controlled medical, multipurpose vocabulary that is subject to further additions. As stated by Cimino *the unit of symbolic processing is the concept - an embodiment of a particular meaning*. Q-Codes can be seen as a medical subject authority list, including medical subject headings, a comprehensive series of mutually exclusive terms. According to guidelines set by Cimino, we have tried to gather a set of non-redundant, shareable, multipurpose, high-quality permanent concepts, in a mono-hierarchical organization, identified by a set of definitions and linked to existing terminologies. This study proposes a system of Knowledge Management (KM) in GP/FM which could potentially fill a major gap in KM of GP/FM. Conceived as a lightweight, multilingual ontology that is fit for new Internet technologies, NLP, and Semantic Web, 3CGP gives the opportunity to unravel GP/FM productivity and establish GP/FM as a professional discipline aiming at an extended range of specific knowledge.

5.1.1 Filling in a major gap in GP/FM and PHC

To the best of our knowledge, there is nothing similar available in GP/FM that has been developed for both human and machine use. There is also not anything of this measure that demonstrates the complexity of GP/FM. Due to the overlap GP/FM with the first line of health service, this tool could also be useful in Primary Care

5.1.2 Paving the way for an ontology in GP/FM and PHC

Though this project took years of work, it acts only as a base from which future researchers may expand upon. As it was designed according to terminological/ontological concepts, is available in OWL and is ready for use with Linked Data. The set of ICPC-2 and Q-Codes is a lightweight ontology; however, because it adapts NLP and automatic and semiautomatic coding it could serve as the basis for the development of a real ontology in GP/FM. The path to a real ontology is still a long time in the making.

5.1.3 Opening the gates for multilingualism in GP/FM and PHC

English has always been the fall-back language of GP/FM. However, family doctors speak to their patients and with one another in their own language, which leads to confusion in translation and varied context of vocabularies. This study has given a potential solution to this issue, by allowing for ICPC-2 to be published in 20 languages and Q-Codes in 8 languages.

Study limitations

5.1.4 A Single-Researcher Study

An important issue to address is that there was a seven-year hiatus in this research, shown by the dates of the conference abstracts analyzed. This was due to an extended illness by the main author. Despite this hiatus, research was eventually able to move forward. Any negative effects resulting from this hiatus may be offset by the fact that only one researcher analyzed the abstracts. Bradley, Curry, and Devers (2007), qualitative data analysis experts, argue that *a single researcher conducting all the coding is*

both sufficient and preferred.[...]. In such cases, the researcher is the instrument; data collection and analysis are so intertwined that they should be integrated in a single person who is the choreographer of his/her own dance[...] However, bias of said researcher could have influence over the collection of data and its analysis.

5.1.5 An empirical move

The Q-Codes form the initial building blocks of classification in the GP/FM field. However, this approach has been filled with the personal experience of the main researcher, which may lead to unintentional biases. One can argue that the qualitative approach to the coding process is both inductive and deductive, an approach sometimes called abductive (Silver and Lewins, 2014). As an empirical document, one has tried to change, fill in the gaps and modify content of classifications using GP/FM publications, pair experience, critiques, and application to real work. MeSH's corresponding descriptors, searching, and indexation exercises on published documents have been also a good way to verify the applicability of the classifications.

5.1.6 Validity and reliability

Another potential issue is the validity of identification and concepts generated. Validity is concerned with *whether a variable measures what it is supposed to measure* (Bollen 1984 cited by Adcock and Collier, 2001). Here, we deal with the identification of concepts in texts. Yet, how can we measure that the same text will generate the same concepts accurately? Ambiguity may pose issue in the execution of this project.

Evaluating *Consistency - Reliability - Interpretive Validity* is referring to whether these Q-Codes could be tested. It was imperative that the Q-Codes could be evaluated through extensive use GP/FM grey literature indexation before being considered a valid construct.

5.1.7 A searcher bias in need of discussion

We are not proposing a standard; however, we are proposing a searcher bias in need of discussion. The main aim of this research is to facilitate the management of information produced by family doctors and to prepare it for further computerized development/reuse. The need to manage GP/FM information in a structured and standardized way must remain a substantial facet of research. The future of the profession is at stake.

It is important to recognize that Q-Codes have been created from a limited number of abstracts. If a concept was not present in the read abstracts, it will have no place in the Q-codes. This emphasizes that the current program is limited to a small number of abstracts within the GP/FM field. Q-Codes would need to integrate much more information to be considered a fully applicable program to the field of GP/FM. Further conferences will contribute new concepts to this, while simultaneously helping GP/FM to evolve. We hope that the structure of this proposed taxonomy will remain enough strong to support the introduction of new items, but it must be taken into account that as more information is added, the basis could potentially not be strong enough to accommodate all.

One issue with the Q-codes ontology involves the unique identifiers. Cimino notes that when building an ontology, there is an *irresistible temptation to make the unique identifier a hierarchical code which reflects the concept's position in the hierarchy*. However, there are inherent disadvantages to using unique identifiers. The first issue, which we have encountered here, is that the coding system runs out of room to grow (Cimino, 1996). This can be due to limited depth, limited breadth, or both of the unique identifier. For instance, when the code has a limited number of positions (digits), the depth of the hierarchy is limited.

Further ontological research is needed to determine whether the two main rules of taxonomic thinking have been respected: completeness (all identified) and exclusivity (a place for each concept) (Ittoo and Bouma, 2013).

5.1.8 Advantages and limits of The Semantic Web

The Q-Codes bases, like all the terminologies edited on HeTOP server, are fit for The Semantic Web. Semantic Web technologies promote common data formats and exchange protocols on the Web, like the Resource Description Framework (RDF), the cited OWL (now available OWL-2) and the query language SPARQL. We have seen that Linked Open Vocabulary (LOV) (<http://lov.okfn.org/dataset/lov/>), a lightweight ontology, differentiates from other ontologies through its characteristics that enable reuse and integration of other vocabularies-i.e., (i) small size, (ii) low formal constraints, (ii) few instances

except for examples, (iii) rich user documentation. Labels, comments, definition, description, etc. are all characteristics of Q-Codes and ICPC-2 on the HeTOP server.

Nevertheless semantic difficulties may arise from the supposed simplicity of the language. The relation *is a* is not an simple relation. Aristotelian logic, which decomposes the proposition into subject and predicate (Younes, 2016), is not sufficient in rendering reality. Following Wittgenstein, the relation *is a* has at least three semantic interpretations. As stated by Wittgenstein (1922) in *Tractatus Logico Philosophicus* (TLP 3.323): *In the language of everyday life it very often happens that the same word signifies in two different ways – and therefore belongs to two different symbols – or that two words, which signify in different ways, are apparently applied in the same way in the proposition. Thus the word “is” appears as the copula, as the sign of equality, and as the expression of existence [.] In the proposition “Green is green” – where the first word is a proper name as the last an adjective – these words have not merely different meanings but they are different symbols.* Language could be more complex than its use in Health Information systems. As stated by Elish and Boyd (2017); *Because computational systems require precise definitions and mathematically sound logics, sociocultural phenomena that are typically nuanced and fuzzy are rendered in coarse ways when implemented into code.* Again, the last word will be given to Wittgenstein (TLP 4.002): *Language disguises the thought. So that from the external form of the clothes one cannot infer the form of the thought.*

6 CONCLUSION

Constructed on the basis of Semantic web technologies, Q-Codes could be considered as a lightweight ontology ready to be used in the semantic web domain, to be extracted in OWL. The multilingual classes of the classification could be individually reached through Unique Resource Identifiers (URIs). Note that each entry gives access to a detailed terminological description, mappings to other terminologies like Babelnet and Dbpedia and to automatic queries on resources like PubMed.

We have created a terminology that highlights the vastness of GP/FM contributions to medical knowledge. We hope, by doing this, to contribute to the recognition of GP/FM as a professional entity within the scientific community that contributes heavily to all fields of medicine. Several questions remain unsolved. Does the current extent of the knowledge base efficiently cover the GP/FM domain? How will this resist the hierarchical structure proposed to the introduction of new themes? Will this system retain enough inter-observers reliability? Nevertheless, given the number of contributions by volunteer translators, such an indexing system seems largely expected by the profession. We hope to transform it into a validated tool for its development.

REFERENCES

- AAFP (2011). *Primary Care definitions – AAFP Policies*. URL: <http://www.aafp.org/about/policies/all/primary-care.html> (visited on 02/07/2016).
- Adcock, Robert and David Collier (2001). “A Shared Standard for Qualitative and Quantitative Research”. In: *American Political Science Review* 95.3, pp. 529–546. ISSN: 00030554. DOI: 10.1017/S0003055401003100. arXiv:1011.1669v3.
- Allen, Bradley P (2016). “The role of metadata in the second machine age”. In: *Second International Conference Establishment Surveys*. ISBN: 9788578110796. URL: <http://www.amstat.org/meetings/ices/2000/proceedings/S57.pdf>.
- Allen, Justin et al. (2011). “The European Definition Of General Practice / Family Medicine”. In: URL: <http://www.woncaeurope.org>.
- Angell, Marcia (2017). “Drug Companies & Doctors: A Story of Corruption”. In: Bachman, CW (1969). “Data structure diagrams”. In: *SIGMIS Newsletter* 1.2, pp. 4–10. DOI: 10.1145/1017466.1017467. URL: <http://dl.acm.org/citation.cfm?id=1017467>.
- Bentzen N.(ed) (2003). *WONCA Dictionary of general/family practice*. Ed. by Niels Bentzen. Maanedsskr. Copenhagen. URL: <http://www.ph3c.org/PH3C/docs/27/000092/0000052.pdf>.
- Bradley, Elizabeth H, Leslie A Curry, and Kelly J Devers (2007). “Qualitative data analysis for health services research: developing taxonomy, themes, and theory.” In: *Health services research* 4, pp. 1758– 72. ISSN: 0017-9124. DOI: 10.1111/j.1475-6773.2006.00684.x.
- Britt, H et al. (2003). “Bettering the Evaluation And Care of Health 2001-2002 (summary of results)”. In: *Australian Family Physician* 32.1/2, pp. 59–63. URL: <http://www.racgp.org.au/document.asp?id=8998>.
- Britt, Helena et al. (2016). *A decade of Australian general practice activity 2006–07 to 2015–16. Bettering the Evaluation and Care of Health (BEACH)*. Report. Family Medicine Research Centre. Sydney School of Public Health. University of Sydney. URL: https://ses.library.usyd.edu.au/bitstream/2123/15482/5/9781743325162__ONLINE.pdf.

- Buono, Nicola et al. (2013). "40 years of biannual family medicine research meetings – The European General Practice Research Network (EGPRN)". EN. In: *Scandinavian Journal of Primary Health Care*. URL: <http://www.tandfonline.com/doi/full/10.3109/02813432.2013.847594>.
- Cabot, Chloé et al. (2017a). "Evaluation of the Terminology Coverage in the French Corpus LiSSa." In: *Studies in health technology and informatics* 235, pp. 126–130.
- Cardillo, Elena (2015). "Mapping between international medical terminologies to SHN Work Package 3". In: *SemanticHealthNet*. Chap. Deliverable 3.3, 18p.
- Carey, Iain M et al. (2004). "Developing a large electronic primary care database (Doctors' Independent Network) for research". In: *International Journal of Medical Informatics* 5, pp. 443–453. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2004.02.002.
- Casado Vicente, Verónica (2012). *Tratado de medicina de familia y comunitaria*. Ed. by Verónica Casado Vicente. Médica Panamericana, p. 2563. ISBN: 8498355850.
- Cavadas, L F, T Villanueva, and J Gervas (2010). "General practice innovation: a Portuguese virtual conference". In: *Med Educ* 44.5, pp. 514–515. ISSN: 0308-0110. DOI: 10.1111/j.1365-2923.2010.03649.x.
- Charlton, Rachel A et al. (2010). "Identifying major congenital malformations in the UK General Practice Research Database (GPRD)". In: *Drug Safety: An International Journal of Medical Toxicology and Drug Experience* 9, pp. 741–750. ISSN: 0114-5916. DOI: 10.2165/11536820-000000000-00000.
- Chinitz, David P and Victor G Rodwin (2014). "Perspective On Health Policy and Management (HPAM): mind the theory-policy-practice gap". In: *International Journal of Health Policy and Management* 3.x, pp. 1–3. DOI: 10.15171/ijhpm.2014.122.
- Cimino, JJ (1996). "Review paper: codingsystems in health care." In: *Methods of information in medicine* 35.4-5, pp. 273–84. ISSN: 0026-1270. URL: <http://www.ncbi.nlm.nih.gov/pubmed/9019091>. — (1998). "Desiderata for controlled medical vocabularies in the twenty-first century". In: *Methods of Information in Medicine* 37.4-5, pp. 394–403.
- Colliers, Annelies et al. (2016). "Improving Care And Research Electronic Data Trust Antwerp (iCARE- data): a research database of linked data on out-of-hours primary care". In: *BMC Research Notes* 9.1, p. 259. ISSN: 1756-0500. DOI: 10.1186/s13104-016-2055-x.
- Davis, D. A. (2004). "CME and the pharmaceutical industry: two worlds, three views, four steps". In: *CMAJ*. Vol. 171, pp. 149–50. ISBN: 0820-3946 (Print)1488-2329 (Electronic). DOI: 10.1503/cmaj.1040361.
- de Oliveira Tavares, A., Santiago, L. M., Jamouille, M., Rosendo, I., & Simões, J. A. (2019). Context and practice of the final work of pre graduate student in General Practice / Family Medicine. Use of 3GCP in Integrated Master's of Medicine at the University of Coimbra. Family Medicine and Primary Care Review (accepted). Retrieved from <https://orbi.uliege.be/handle/2268/226901>
- Dixon, Brian E, Atif Zafar, and Julie J McGowan (2007). "Development of a taxonomy for health information technology." In: *Studies in health technology and informatics* 129.Pt 1, pp. 616–620. ISSN: 0926-9630.
- Druais, PL et al. (2009). *Médecine générale*. Ed. by D Pouchain. Masson SA, p. 460.
- Elish, M. C. and Danah Boyd (2017). "Situating methods in the magic of Big Data and AI". In: *Communication Monographs*, pp. 1–24. ISSN: 0363-7751. DOI: 10.1080/03637751.2017.1375130.
- Glaser, Barney G. and Anselm L. Strauss (1999). *Discovery of Grounded Theory: Strategies for Qualitative Research - - Livres*. AldineTransaction, p. 284. ISBN: 0202302601.
- Goggi, S. et al. (2015). "A semantic engine for grey literature retrieval in the oceanography domain". In: *Seventeenth International Conference on Grey Literature*. Amsterdam, pp. 76–77. URL: http://greyguide.isti.cnr.it/wp-content/uploads/2017/04/GL17{_}Program{_}Book-min.pdf.
- González-González, A I et al. (2007). "Information Needs and Information-Seeking Behavior of Primary Care Physicians". In: *Ann Fam Med* 4, pp. 345–352. ISSN: 1544-1709 (Print)1544-1717 (Electronic). DOI:10.1370/afm.681.
- Greenway, Tyler and Joseph S Ross (2017). "US drug marketing: how does promotion correspond with health value?" In: DOI: <https://doi.org/10.1136/bmj.j1855>.
- GreyNet (2014). *Pisa Declaration on Policy Development for Grey Literature Resources*. URL: <http://greyguiderep.isti.cnr.it/Pisadeclapdf/Pisa-Declaration-May-2014.pdf> (visited on 06/16/2017).
- Grosjean, Julien et al. (2012). "Teaching medicine with a terminology/ontology portal." eng. In: *Studies in health technology and informatics* 180, pp. 949–53. URL: <http://europepmc.org/abstract/MED/22874333>.
- Gusso, Gustavo and José Mauro Ceratti Lopes (2012). *Tratado de Medicina de Família e Comunidade: 2 Volumes: Princípios, Formação e Prática*, p. 2180. ISBN: 8536327979.
- Gómez-Pérez, A., M. Fernández-López, and O. Corcho (2003). "Ontological Engineering and the Semantic Web". URL: http://www.exa.unicen.edu.ar/escuelapav/cursos/corcho/01_introduction.pdf.
- Helman, Cecil G (2008). *Medical Anthropology*. Ashgate, p. 580. ISBN: 978-0-7546-2655-8.
- Heyrman, J (ed.) (2005). *EURACT Educational Agenda, European Academy of Teachers in General Practice EURACT*, Leuven.
- Hubbard, Derek (2008). *How to Find Clinical Information Quickly at the Point of Care*. Vol. 15. 6. American Academy of Family Physicians, p. 23.
- Hummers-pradier, Eva (2007). "Which Abstracts Do Get Published ? – Output Of German Gp Research 1999-2003". In: *Wonca Europe Paris 2007*.

- Ittoo, Ashwin and Gosse Bouma (2013). "Term extraction from sparse, ungrammatical domain-specific documents". In: *Expert Systems with Applications* 40, pp. 2530–2540.
- James, Jack E (2016). "Free-to-publish, free-to-read, or both? Cost, equality of access, and integrity in science publishing". In: *Journal of the Association for Information Science and Technology* 68.6, pp. 1584–1589. ISSN: 2330-1643. DOI: 10.1002/asi.23757.
- Jamoulle, M et al. (2014). "Mapping French terms in a Belgian guideline on heart failure to international classifications and nomenclatures: the devil is in the detail". eng. In: *Inform Prim Care* 21.4, pp. 189–198. DOI: 10.14236/jhi.v21i4.66.
- Jamoulle, Marc (2015). "Quaternary prevention, an answer of family doctors to overmedicalization". In: *International Journal of Health Policy and Management* 4.2, pp. 61–64. ISSN: 2322-5939. DOI: 10.15171/ijhpm.2015.24. URL: http://ijhpm.com/article{_}2950{_}0.html.
- Jamoulle, Marc, Julien Grosjean, and Stefan Darmoni (2017). "Access to multilingual individual rubrics in URI format for ICPC-2 and the Q-Codes". In: URL: <http://orbi.ulg.ac.be/handle/2268/211268>.
- Jamoulle, Marc et al. (2015). "Semantic Web and the Future of Health Care Data in Family Practice". In: *Merit Research Journal of Medicine and Medical Sciences* 3.12, pp. 586–594. URL: <http://orbi.ulg.ac.be/handle/2268/189292>.
- Jamoulle, M. (2017). Towards a system of concepts for Family Medicine . Multilingual indexing in General Practice / Family Medicine in the era of Semantic Web. Doctoral thesis. University of Liege, Belgium. Retrieved from <http://hdl.handle.net/2268/216132>
- Jamoulle, Marc et al. (2017a). "A terminology in General Practice / Family Medicine to represent non-clinical aspects for various usages : the Q-Codes". In: *Medical Informatics Europe (MIE2017) Informatics for Health 2017 / April*, pp. 1–5. URL: <http://orbi.ulg.ac.be/handle/2268/206527>.
- Jamoulle, Marc et al. (2017b). "Analysis of definitions of General Practice/Family Medicine and Primary Health Care". In: *British Journal of General Practice - Open*, 050 ISSN: 0960-1643. URL: <http://orbi.ulg.ac.be/handle/2268/210049>.
- Janamian, T et al. (2016). "Quality tools and resources to support organisational improvement integral to high-quality primary care: a systematic review of published and grey literature". In: *Med J Aust* 204.7 Suppl, S22–8. ISSN: 0025-729x.
- Jonquet, Clement et al. (2016). "SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique". In: *16e Journées Francophones d'Informatique Médicale(JFIM)* June.
- Khan, Nada F, Sian E Harrison, and Peter W Rose (2010). "Validity of diagnostic coding within the General Practice Research Database: a systematic review". In: 60.572, e128–e136. ISSN: 0960-1643. DOI:10.3399/bjgp10X483562.
- Kochen, Michael M. (2012). *Allgemeinmedizin und Familienmedizin*. Thieme; Auflage: 4., vollständig überarbeitete und erweiterte Auflage, p. 652. ISBN: 3131413840.
- Kruschinski, C., Lange, M., Lionis, C., van Weel, C., & Hummers-Pradier, E. (2010). Themes and methods of research presented at European General Practice Research Network conferences. *Family Practice*, 27(4), 459–67. <https://doi.org/10.1093/fampra/cmz023>
- Lakhani, Mayur K. (2003). *A Celebration of General Practice*. Radcliffe Publishing, p. 203. ISBN: 1857759230.
- Lawrence, Amanda et al. (2014). "Where is the evidence: realising the value of grey literature for public policy and practice". In: *Australian Policy Online*. DOI: 10.4225/50/5580B1E02DAF9. URL: <http://apo.org.au/node/42299>.
- Lelong, R et al. (2016). "Semantic Search Engine to Query into Electronic Health Records with a Multiple- Layer Query Language". In: *MEDIR workshop*. URL: http://medir2016.imag.fr/data/MEDIR_2016_paper_8.pdf.
- Liang, S. F. et al. (2014). "Semi Automated Transformation to OWL Formatted Files as an Approach to Data Integration". In: *Methods of Information in Medicine* 54.1, pp. 32–40. ISSN: 0026-1270. DOI: 10.3414/ME13-02-0029.
- Lin, Jennifer and Carly Strasser (2014). "Recommendations for the role of publishers in access to data." In: *PLoS biology* 12.10, e1001975. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001975.
- Lowe, H J and G O Barnett (1994). "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches." In: *JAMA* 271.14, pp. 1103–8. URL: <http://www.ncbi.nlm.nih.gov/pubmed/8151853>.
- Madkour, Mohcine, Driss Benhaddou, and Cui Tao (2016). "Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain". In: *Computer Methods and Programs in Biomedicine*, pp. 52–68. DOI: 10.1016/j.cmpb.2016.02.007.
- Mahood, Quenby, Dwayne Van Eerd, and Emma Irvin (2014). "Searching for grey literature for systematic reviews: challenges and benefits". In: *Research synthesis methods* 5.3, pp. 221–34. DOI: 10.1002/jrsm.1106.
- McGuinness, DL and Frank van Harmelen (2004). *Web Ontology Language*. URL: <http://www.w3.org/TR/owl-features/>.
- McIntyre, Ellen et al. (2016). "The contribution of a knowledge exchange organisation in primary healthcare." In: *Australian family physician* 45.9, pp. 684–7. ISSN: 0300-8495. URL: <http://www.ncbi.nlm.nih.gov/pubmed/27606374>.
- McKenzie, Lindsay (2017). "Sci-Hub's cache of pirated papers is so big, subscription journals are doomed, data analyst suggests". In: *Science*. ISSN: 0036-8075. DOI: 10.1126/science.aan7164.
- McWhinney, Ian R. (1997). *A Textbook of Family Medicine*. Oxford University Press, p. 448. ISBN:019511518X.
- Moynihan, R (2003). "Who pays for the pizza? Redefining the relationships between doctors and drug companies. 1: entanglement". In: *Bmj* 326.7400, pp. 1189–1192. ISSN: 0959-535x. DOI: 10.1136/bmj.326.7400.1189.
- Moynihan, R and L Bero (2017). "Toward a Healthier Patient Voice: More Independence, Less Industry Funding". In: *JAMA Intern Med* 177.3, pp. 350–351. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.9179.

- Murtagh, John (2011). *John Murtagh's General Practice*. McGraw-Hill Medical Publishing Division, p. 1508. ISBN: 0070285381.
- PAHO Bireme Sao Paulo (2016). *Virtual Health Library Search Portal of the Regional library of Medicine*.
URL: <http://regional.bvsalud.org/http://www.paho.org/bireme/>.
- Quan, Wei, Bikun Chen, and Fei Shu (2017). "Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016)". In: *ArXiv e-prints*. arXiv: 1707.01162. URL: <http://arxiv.org/abs/1707.01162>.
- Resnick, M. P., Santana, F., de Araujo Novaes, M., Shamenek, F. S., Frieden, L., & Iyengar, M. S. (2013). Representing second opinion requests from primary care within the Brazilian tele-health program: international classification of primary care, second edition. *Studies in Health Technology and Informatics*, 192, 1190. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23920964>
- Rigaux, Sébastien (2015). "Classification automatisée de résumés médicaux – Belgique 2015 [Multi-Label Text Classification of Medical Abstracts]". Master thesis.
- Salman Bin Naeem, Salman, Ahmed Shamshad, and Khan Amjid (2013). "Information seeking in primary care: a survey of doctors working in remote government health facilities in Pakistan". In: *Library Philosophy and Practice (e-journal)* Paper 1009. URL: <http://digitalcommons.unl.edu/libphilprac/1009>.
- Schrans, D. et al. (2016). "The search for person-related information in general practice: a qualitative study". In: *Fam Pract* 33.1, pp. 95–9. ISSN: 0263-2136. DOI: 10.1093/fampra/cmz099.
- Shen, Cenyu and Bo-Christer Björk (2015). "'Predatory' open access: a longitudinal study of article volumes and market characteristics". In: *BMC Medicine* 13.1, p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-015-0469-2.
- Silver, Christina and Ann Lewins (2014). *Using Software in Qualitative Research*. SAGE Companion. URL: <https://study.sagepub.com/using-software-in-qualitative-research>.
- Soler, Jean karl, Marc Jamouille, and Peter Schattner (2015). "The International Classification of Primary Care". en. In: *The World Book of Family Medicine – European Edition 2015*. Ljubljana. ISBN: 978-961-281-983-5. URL: <http://orbi.ulg.ac.be/handle/2268/187050>.
- Soler, Jean K. et al. (2012). "The interpretation of the reasons for encounter 'cough' and 'sadness' in four international family medicine populations". In: *Informatics in Primary Care* 20, pp. 25–39. ISSN: 14760320.
- Sutton, Stuart A (2007). "Tutorial 1: Basic Semantics". In: August. URL: <http://dublincore.org/resources/training/dc-2007/T1-BasicSemantics.pdf>.
- Swan, Alma. (2012). *Policy guidelines for the development and promotion of open access*. United Nations Educational, Scientific, and Cultural Organization, p. 76. ISBN: 9789230010522.
- Tabatabaei-Malazy, O, S Nedjat, and R Majdzadeh (2012). "Which information resources are used by general practitioners for updating knowledge regarding diabetes?" In: *Arch Iran Med* 4, pp. 223–227. ISSN: 1029-2977. DOI: 012154/aim.0010.
- Thompson, C. A. et al. (2014). "Patient and provider characteristics associated with colorectal, breast, and cervical cancer screening among Asian Americans". In: *Cancer Epidemiol Biomarkers Prev*. Vol. 23. United States: (c)2014 American Association for Cancer Research., pp. 2208–17. ISBN: 1538-7755 (Electronic)1055-9965 (Linking). DOI: 10.1158/1055-9965.epi-14-0487.
- Van Royen, Paul et al. (2010). "Are presentations of abstracts at EGPRN meetings followed by publication?" EN. In: *The European journal of general practice* 16.2, pp. 100–5. ISSN: 1751-1402. DOI:10.3109/13814788.2010.482582.
- VanNieuwenborg, L et al. (2016). "Continuing medical education for general practitioners: a practice format". In: *Postgrad Med J* 92.1086, pp. 217–222. ISSN: 0032-5473. DOI: 10.1136/postgradmedj-2015-133662.
- Vanopstal, Klaar et al. (2011). "Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot." In: *Journal of medical systems* 35.4, pp. 527–43. ISSN: 0148-5598. DOI: 10.1007/s10916-009-9389-z.
- Veuillotte, I et al. (2015). "General practice and the Internet revolution. Use of an Internet social network to communicate information on prevention in France". In: *Health Informatics J* 1, pp. 3–9. ISSN: 1460-4582. DOI: 10.1177/1460458213494905.
- Wittgenstein, Ludwig (1922). *Tractatus logico-philosophicus*. London: Kegan Paul, p. 108. ISBN: 1602064512.
URL: <http://people.umass.edu/klement/tlp/>.
- WONCA (1987). *ICPC: International Classification of Primary Care*. Ed. by H Lamberts and M Wood. Oxford: Oxford University Press. ISBN: 0 19 261633 1.
- WONCA (2005). *ICPC-2-R: International Classification of Primary Care (Oxford Medical Publications)*. Oxford University Press, USA, p. 204. ISBN: 019262802X.
- Younes, Nora (2016). *Introduction à Wittgenstein*. Paris: Editions La Découverte, p.1

Library, Information Science & Technology AbstractsTM with Full Text

Available via EBSCOhost[®]

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

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

Comprehensive content includes:

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

Subject coverage includes:

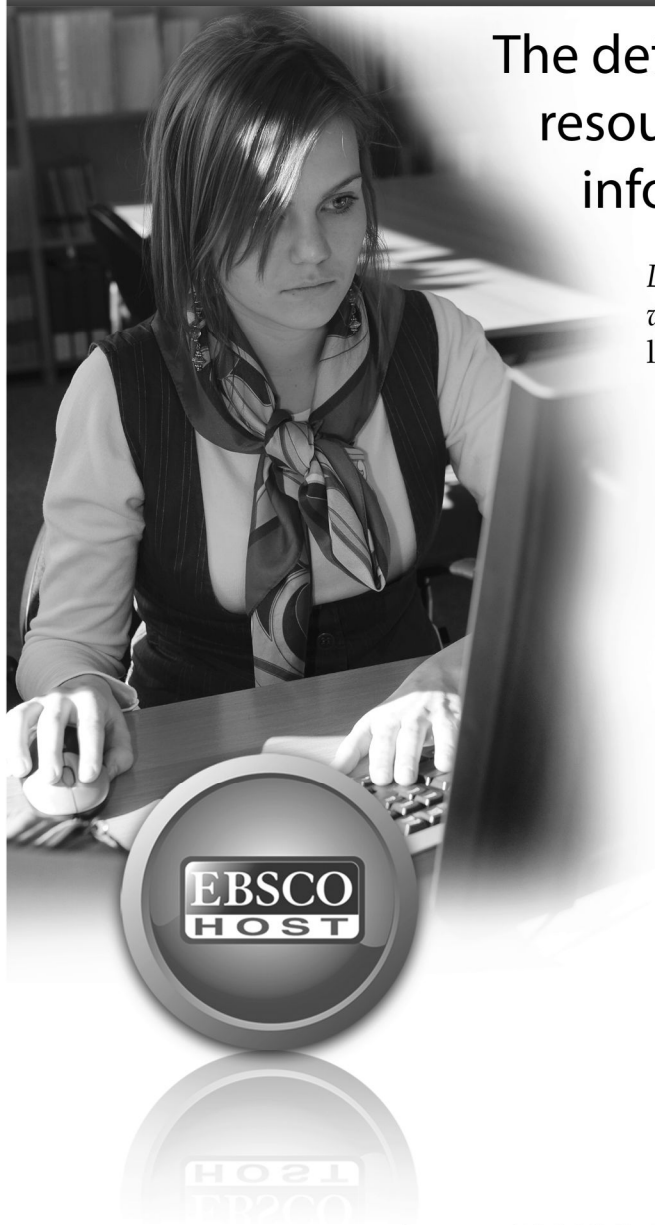
- Bibliometrics
- Cataloging
- Classification
- Information Management
- Librarianship
- Online Information Retrieval
- And much more...

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

Phone: 800.653.2726

Email: request@ebscohost.com

www.ebscohost.com



Apps & Codes: Making profiles for fluid publishing contents^{*}

Flavia Cancedda <https://orcid.org/0000-0001-7564-495X>

Luisa De Biagi (Italy)

Abstract

Apps are one of the most used digital publishing tools and convey intellectual contents for innumerable functionalities. App programs present technical characteristics different from each other, depending on the intellectual content for which they are profiled and made available. In publishing field, apps that convey traditional editorial products disclose the maximum of their innovative potential in representing the interface of continuously updating products (such as newsletters, magazines, newspapers, guides and tourist maps, blogs, open/e-gov/research data systems and related data-bases, clinical trials, forums and websites with specific matter or content, etc.). If apps are the interface software – thus, the communication interface - of publishing products, could it be possible to make them identifiable (and manageable) through the same bibliographic codes used for the corresponding traditional publishing products (e.g., ISBN, ISSN, DOI, etc.)? We think that when apps show and preserve the essential bibliographic identifying data, the answer could be no less than positive.

The goal of this study is to show, through a comparative analysis of National and European case-studies and best practices examples, how the potential public usefulness (not only commercial but also informative) of a bibliographic identification for apps is similar to the usefulness recognized in case of traditional publishing products, opening new scenarios and results also for grey products: a new traceability skill could be established for apps containing permanent references, traditionally considered necessary for the identification of editorial objects (title, publisher, year, updating mode or frequency, etc.); traceability through numerical codes would allow also easier dissemination, marketing, indexing processes by search engines, portals and sales store, up to make indexing tools for bibliographic services and librarians more specific and professional.

Apps&codes

Apps are one of the most used digital publishing tools... and convey intellectual contents for innumerable functionalities. We would like to start from this statement – basic and almost obvious - to introduce a subject that is increasingly focusing in publishing environment. Preventively, it's better to emphasize that we are here conceptually moving within publishing and media area or environment: we do not mean to refer to apps outside of this context or to apps providing non-editorial services (e.g.: games, advertising or promotional tools, information on transport timetables, personal physical functions monitoring, weather forecast etc.), neither to apps whose function is only to facilitate the connections of mobile devices to traditional websites. Moving within the publishing industry, we hereby present the knowledge and therefore the practical experience of the ISSN Italian Center (member of the International ISSN Network¹): for its function, it occupies a privileged point of observation on the panorama of publishing innovations.

ISSN National Centers deal with registering and identifying serial publications (magazines, newspapers, monographic series, websites or updating databases, and so on...). Centers operate in a field of activity undoubtedly relevant to librarianship, also strongly focused on the *techniques of bibliographic identification* (more than on bibliographic description, subject indexing or other specific library activities). ISSN Centers are also particularly involved in the *application policies* of those techniques, keeping constantly in touch with the changing international phenomena of publishing. For many years, there were questions in identifying digital (online) editorial/intellectual products with features different from those typical of a magazine, a book, a monographic series published or marketed online. A traditional online magazine, or book, or monographic series keeps several editorial indicators – a kind of signposting, the paratextual areas - borrowed from paper publishing: a homepage acts as a cover even when there is not digital cover; one or more webpages dedicated to presentation, contacts, acting as a colophon; another web page dedicated to the index, etc. It is not particularly difficult to assimilate

^{*} First published in the GL19 Conference Proceedings, February 2018.

¹ The International Standard Serial Number Network is composed by 89 national Centers, and coordinated by the International Center in Paris, www.issn.org.

cataloging and identification of these products to that used for paper products: the physical support difference makes it necessary to add or modify some descriptive details, but does not affect the possibility of circumscribing, and thus identifying these digital products, for all the purposes for which an identification is useful.

But now we are not talking about an online magazine, which can also be searched by smartphones or about an e-book, which can also be read in a tablet. Instead, we would like to point out that the apps we frequently download on our devices are something more and different from a "traditional" digital editorial product, even when they aim to convey a publishing product that has a pretty traditional source. Apps are born as programs, not as "contents"; they arise as a tool or vector, and not as "editorial form". In short, they allow you to use intellectual content on the latest generation of devices: the programs that allow this type of use (via tablet, e-book reader, iphone, smartphone...), with technological and usage parameters defined by the type of device involved, are synthetically called "apps". There are, in fact, apps with a traditional interface, which allow them to be optimally visualized on those devices but do not significantly affect the editorial product they deliver: many e-books apps have these features. But there are also "less-traditional" apps (those we are currently reflecting on) for *app-native editorial products* where the intellectual or informational content has been profiled - we could say: editorially imprinted, constructed - to exploit the technological potential of the app. Our work experience leads us to meet more and more traditional editorial products that are substantially re-designed to be used through apps and to exploit the undoubted technological advantages of apps: e.g. newspapers or weekly magazines, scientific databases, thesauri, interactive manuals, guides and tourist maps, personal directories or address books, commercial catalogs for sales, newsletters focused about specific events (as fairs, meetings, exhibitions...), clinical case-studies directories, multiplatform related to info services etc. Many of these types - and others - start from a traditional editorial idea, but their content is assembled and reassembled through the cooperation of different software components for managing interactions between the publisher, other related content to which the publisher decides to allow access, and the operations of interactive use by the end-user.

Especially regarding the continuous updating and the connectivity with other related digital products or environments, the technology under the app is so incisive on the intellectual content that it is legitimate to ask: are these "editorial" products actually so? and, if so, are they still identifiable as such (i.e., by same tools and methodologies applied to more traditional products)?

If the answer were "no" we should admit that we are faced software: that is, we are facing technological tools, probably intended to change in a short time, as well as all many other technologies: software would not fall into our field of publishing experts; it would not be useful to discuss whether and how to identify software using bibliographic methods.

However, we believe that the answer may be "yes": publishing products transmitted through apps are editorial products, and it is possible - and useful - to describe/identify them using bibliographic methods and tools. From our point of view, we can use bibliographic identification codes (ISBN, ISSN, DOI etc.) to uniquely identify these products, provided these products present the minimum bibliographic characteristics through which they can be identified.

According to the traditional cataloging praxis and bibliographic descriptive literature, the minimum characteristics to identify with sufficient certainty an editorial product are: the product has a title (words, phrases, characters or character groups on the product, intent on identifying it; the title has to be detectable and constant even in following updates or serial outputs of the same product); the product shows a copyright or an editorial responsibility; the product has one or more dates of publication/release/update dates... and so on. If these bibliographic indicators can easily be found in the editorial product delivered via app - e.g. an interactive tour guide that helps us move around in a city, find places and services for us, memorize our preferences, connect us to other information sources or allow chat and messaging - we are faced with an editorial product recognizable and bibliographically identifiable through traditional tools as international identification codes: an interactive tourist guide can be identified through ISBN and also through DOI; and through ISSN too, if the guide explicitly declares its future updates (i.e. if it assumes the characteristics of a serial publication).

We have just mentioned three international reference systems for bibliographic identification, ISBN, ISSN, and DOI: the ISO standards underlying these systems indicate their very extensive application

without restrictions concerning digital environments. For example, the ISBN Manual, par. 6.1 / 6.2 explicitly states: "*Where a publication is available electronically (e.g., an e-book, e-book app, CD-ROM or publication available on the Internet), it will qualify for an ISBN provided that it contains text and is available to the public ... An ISBN may be used to identify a specific software product that is intended for educational and / or instructional purposes, such as a computer-based training product, provided that it is neither customizable nor requires data in order to function*"². The ISO 3297 standard which regulates the ISSN system provides a very large scenario: in fact, ISSN code can be assigned to continuous (=ongoing) integrating resources: "*Continuing resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole*"³.

More specifically and explicitly in the DOI Foundation's *Factsheets*⁴, several definitions and statements confirm that DOI can be easily applied to editorial apps as well:

"DOI® is applicable to any object (= any entity or thing: physical, digital, or abstract; resources, parties, licenses, etc.);

- ▶ *is a digital Identifier of an object = network actionable identifier ("click on it and do something")*
- ▶ *initial focus on entities was documents/media e.g., articles, data sets; now moving into parties, licenses and other sectors*
DOI® provides an actionable, interoperable, persistent link
- ▶ *Actionable – through use of identifier syntax and network resolution mechanism (Handle System®)*
- ▶ *Persistent – through combination of supporting improved handle infrastructure (registry, proxy, etc.) and social infrastructure (obligations by Registration Agencies)*
- ▶ *Interoperable – through use of a data model providing semantic interoperability and grouping mechanisms"*.

After this quick overview on international bibliographic identification codes/tools - and therefore on the feasibility of international identification of publishing apps – two questions are necessary, which will affect the future of bibliographic treatment activities by international identification Agencies:

Question 1): why an app should be bibliographically identified (i.e.: why to assign an ID code such as ISBN or ISSN)?

Question 2): ... and, once answered to the previous question, is it practically advantageous to identify an app by a numeric code?

Let us briefly answer the questions.

- 1) The bibliographic identification of an app - that has been explicitly requested by some publishers - means allowing the publisher/distributor/provider to recognize their product unequivocally for all bibliographic and commercial transactions, just like for a traditional editorial product.

Of course, the unique identification can also be used as a prerequisite for attesting the validity and/or the level of editorial quality or accuracy (quite similar as in the field of scientific publishing when an editorial product to be considered for public funding should be identified by one of the existing international bibliographic codes).

- 2) But why should a publisher, practically, want to identify an app (in the absence of legal obligations or common and shared media practices that make it necessary or at least very appropriate)? Why, for example, did a publisher ask us to identify with ISSN code a travel guide that could be downloaded on smartphone, continuously updated and interactively managed through apps? What would be a bibliographic identification code for, since an interactive app would improbably have registered into a library catalog?

The reasons driving a publisher are connected to commercial and digital distribution procedures: therefore, an international identification Agency would not identify an app for library purposes or others purposes related to dissemination at cultural institutions such as libraries or documentation centers, but because the unique identifier code given to the app would be used as fil-rouge (or maybe Ariadne's

² *ISBN User's Manual*, London, International ISBN Agency, 2012 (International ed.; sixth ed.), p. 13 <<https://www.isbn-international.org/sites/default/files/ISBN%20Manual%202012%20-corr.pdf>>

³ *ISO 3297:2017, Information and documentation – International standard serial number*, Geneva, International Organization for Standardization, 2017, par. 3.3 (partially available at <https://www.iso.org/obp/ui/#iso:std:iso:3297:ed-5:v1:en>).

⁴ <https://www.doi.org/factsheets/DOIKeyFacts.html>

thread) for each business publishing procedure: connecting first release and following releases of the product/app from the IT company to the publisher; delivering the product by publisher to providers and different sales platforms; tracking sale or release of the app and its updates for each buyer or user, and for each connected device; facilitating credits access management and related interactive services; recognizing different service levels according to purchasing/subscription levels; managing subscription procedures, subsequent releases, "patches"; accounting the number of downloading; identifying the commercial product/app for tax purposes, for distribution of revenues in connection with the different industrial or intellectual property rights quotes; etc.

Consequently, considering the app as commercial object, the software and the conveyed intellectual content become much easier to manage by the publisher (which is not an IT company, and does not normally have all technological skills) if he use a single identifier code - better one of the bibliographic codes already known - that accompanies the product through all transactions.

The publisher, as well as most of his customers, keeps a "conservative side" in his mind: he certainly appreciates he can act on technologically innovative elusive products with usual and well-known tested tools (the identifier codes). He appreciates much more if that procedure allows him to incardinate without trauma his innovative app within his usual editorial catalog: there, he will incorporate new apps alongside traditional products, transmitting to customers the idea that the purchase or the enjoyment of both will not lead to tedious changes of habit or too awkward access procedures.

So, for the publisher it's very convenient - both for practical and promotional reasons - that apps and books, apps and magazine, apps and articles give the idea they are commercially the same things, so they are managed by users with the same procedures. The use of identified, approved, and standardized bibliographic codes - such as ISBN, ISSN, DOI, and so on - presents for the media several benefits: the codes are already well-known and recognizable (we must remember that distinguishing the type of code means recognizing the type of publishing product); they are internationally reliable and guaranteed as they are maintained by institutionally solid Agencies; the codes have transparent, consolidated and quite easy acquisition procedures; have a fairly low cost (or even no cost at all); they authenticate the identified product, and "welcoming" it in their sphere of competence they sanction identification, nature and, in some respects, cultural profile as well.

For all these reasons, we believe that traditional identifying systems - far from being an obstacle - can be useful for authors and publishers who want to entrust intellectual and informational content to apps, ensuring it is capillary distributed: technologically advanced and more liquid apps, increasingly popular for their plain use and eye-catching graphic presentation - but even more for easy and immediate recognizability.

Software codes, releases, software repositories and DOI interactive use: applications and trends

ISBN and apps

Regarding ISBN, the Italian national Centre provided by AIE (Italian Publisher Association)⁵, reported us that it's impossible to know to which product and typology the code requested will be assigned. The ISBN user/applicant gets a block (lot) of ISBN codes to be applied in publications, usually books or e-books.

For ISBN Italian Centre, surely app can't be identified with ISBN standard... but it's equally possible to identify an educational and didactic software with its user manual and technical instructions if the manual is fundamental for making the software work and its usefulness is strictly conditioned to the match with the software itself.

A new development has been done with OCR applied to ISBN.

ISBN Scan is an ISBN (International Standard Book Number) reader application available on Google Play. ISBN "RealCodeScan" engine can be easily integrated into various smart-phone applications, including iOS systems. This app can read not only the ISBN bar-code but also the ISBN number itself by the OCR (Optical Character Recognition) powered functions.

There is no even need to press the shutter when reading because ISBN Scan reads snap shot images reflected in the camera just by simply waving the device. It's capable to read the ISBN at high accuracy

⁵ <http://www.aie.it/>

with super-fast speed, “modifying the camera configuration (in-app) for BEST Reading performance”.

One of the main benefits of this app is also the possibility of getting and viewing immediately the detailed information sheet of the book (author, title, publisher, etc.), as well as share them via SNS (Social Network Services) or e-mail.

A similar app is ‘My Library’: a simple application helping users to manage a sort of private library at books.google.com. ‘My Library’ is a totally free app (with no advertising) to keep track of all personal books and wishlist books.

In ‘My library’ it’s possible to add a book searching its ISBN code or even scanning its barcode. In those cases informations (author, title, publisher, publication date etc.) are given from Amazon, Google Books and Open Library.

My Library’s main features are⁶:

- possibility of using phone's camera to scan a barcode
- adding /removing books from books.google.com shelves
- downloading pdf of books in public domain
- creating a personal shelf with personalized recommendations/alerts
- support tablet devices.

Though, there are some criticalities reported by users of this app, starting from the impossibility of cataloguing books by matter or a non-complete recognition of all ISBN codes. As ISBN scan ‘My library’ – runs both on Android (Google play) and iOS (Apple I-Tunes)⁷.

Finally **The US DOI Agency is experimenting a new** qr-encoded Crossref DOI, inspired by Google recent promotion of QR codes: it’s possible to generate a QR Code for any given Crossref DOI (even media gadgets)⁸

Making code citable: Zenodo and GitHub

For Open Science it is important to cite the software used in research studies. It ought to be cited any software making a significant impact. Modern research relies constantly on data analysis and the main mission might be to preserve and cite software in a permanent/sustainable, identifiable and simple way. Innovative Digital repositories like Zenodo can really help us reaching this goal.

DOIs are persistent identifiers obtained only by an agency committed to maintain a reliable level of consistency and preservation of a digital resource. So, as a digital repository, Zenodo registers DOIs through DataCite and preserves these submissions using trusted foundation of CERN’s data centre, alongside the LHC’s 100PB Big Data store (biggest scientific dataset worldwide). The code preserved in Zenodo will be accessible in the future to supply long-term digital storage and preservation. The DOIs will function as perpetual links to the resources. DOI based citations remain even if URL and protocol change, because DOI currently direct to URIs (Uniform Resource Identifier). DOIs has also search engines and indexing services, to support software usage through multiple citations.

Long-term digital storage and preservation. Although it is possible to identify softwares uploaded to platforms like GitHub⁹, these platforms do not issue DOIs and there is not a perpetual guarantee of access to older software: we might have to face with preservation and versioning issues problems (e.g. about which version of the code is being referenced and if is it still available on the hosting website). That’s why it’s important submitting the software/code in a digital repository like Zenodo. Born within the OpenAIRE project Zenodo is a free, open-access research repository¹⁰ which profits from and contributes to provide CERN important initiatives as Open Data services (CERN opendata portal)¹¹.

⁶This app is only a book tracking app it’s not a book reader app

⁷<https://itunes.apple.com/it/app/mylibrary/id1141437096?mt=8>

⁸<https://www.crossref.org/display-guidelines/>

⁹<https://github.com/github>

¹⁰Based on the Invenio open-source software

¹¹<http://opendata.cern.ch/>

Launched at the CERN Data Centre in May 2013 with a grant from the European Commission, Zenodo has a special commitment to sharing, citing and preserving data and code.

Submitting the code to Zenodo and receiving a DOI has become easier since Zenodo's integration to the platform GitHub has been done. Moreover, if preservation is based on releases, as the software changes each release can be cited with its own DOI, giving precise traceability and awareness of the exact code used in a data analysis. Code releases are both archived and 'published' to become public, so that they can be described with rich metadata, an explanatory abstract and a significant author list. This means that software developers could theoretically skip the journal publication step to announce and make their creation immediately visible.

Zenodo allows also to sign up only one time by choosing the GitHub or the Zenodo account, in order to advance the immediate interlinking. Finally, if a research is funded by an EU grant, it's even possible to directly connect the code to the grant by updating the grant section of the metadata on the Zenodo record.

References

- ▶ ISO 2108:2005, International standard book number (ISBN)
- ▶ ISO 26324:2012, Digital object identifier system [DOI]
- ▶ ISO 17316:2015, International standard link identifier (ISLI)
- ▶ ISO 3297:2017, International standard serial number (ISSN)

Past Internet standards:

- ▶ RFC 1630 (1994), <https://tools.ietf.org/html/rfc1630>
- ▶ RFC 2141 (1997), <https://tools.ietf.org/html/rfc2141>
- ▶ RFC 3986 (2005), <https://tools.ietf.org/html/rfc3986>
- ▶ <https://ec.europa.eu/research/openscience/index.cfm>

Data from:

“Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication¹”

<https://doi.org/10.17026/dans-zy8-fcjlw>

Dominic Farace, GreyNet International, Netherlands

<https://orcid.org/0000-0003-2561-3631>

Project Administration / Dataset Creator / Draft Text

Joachim Schöpfel, University of Lille, France

<https://orcid.org/0000-0002-4000-807X>

Review and Editing

Abstract

The data collected is based on the responses of forty-four authors/researchers, who have presented full-text papers published in the International Conference Series on Grey Literature². The dataset consists of responses to ten questions, two of which are open-ended. Six of the questions are taken from a questionnaire carried out in 2011 within GreyNet’s Enhanced Publications Project. The data contain the number of responses and percentages to each question. The questions deal with the author’s own empirical research data, its availability, the formats in which it appears, and the author’s willingness to archive it and make it openly accessible. Additional questions deal with the respondent’s citation and reference to data, their use of data journals in carrying out search and retrieval, and whether they have (co)authored a data paper or data article.

The data was collected via SurveyMonkey³, where it remains stored along with a copy in .ods format in the DANS EASY Archive⁴. The reuse of the questions and data collected from the study in 2011 is likewise based on the same selection used to define the population of the earlier study. The data collected allows for insight into the changing attitudes and practices within GreyNet’s research community and renders a use case for other research communities.

Keywords: Enhanced Publication, Data Citation, Data Reuse, Attitude Change

Subject Area: Information Science

Methods Applied

● Steps

In producing the dataset, an online questionnaire was drawn-up and implemented using the freeware SurveyMonkey. The questionnaire contains ten queries, two of which are open-ended. Six of the questions are taken from a survey carried out in 2011. The open-ended questions allow for source data related to file formats in which the data is contained as well as the names of the respondents, their gender, email address, and number of times they published in the GL-Conference Series.

● Sampling Strategy

The population of the survey was selected from among 115 first authors in the International Conference Series on Grey Literature from 2012 (GL14) to 2017 (GL19). Also included were the first authors to the 2018 (GL20 Conference-in-spec) as well as respondents to GreyNet’s 2011 Survey on Enhanced Publications. Once the survey population was reduced, either because an author’s email address was currently unavailable, the author had retired or had since moved to another field, the questionnaire was then sent to the remaining 94 authors/researchers via personalized emails. The final results of the study rest on the responses of 44 survey respondents, which accounts for a 46,8% response rate.

1 st Authors 2012-2018	Survey Recipients	Survey Respondents	Survey Results %
115	94	44	46,8%

● Quality Control

While there was no specific control carried out on the data, the willingness of the respondents to provide their name and email address affords some level of guarantee. Even for those respondents, who chose to remain anonymous, the fact that the survey population was drawn from a designated research community is considered a quality indicator.

Dataset Description

File name:	Survey Results 2018.ods
Format name and version:	ODS 2018
Creation dates:	from 2018-05-18 to 2018-06-15
Language:	English
License:	CC0 Waiver - no rights reserved
Archive name:	DANS EASY Archive
Publication date:	2018-09-06

Potential Reuse of the Data

While the data collected in this study is openly accessible and bears a persistent identifier, its potential for reuse depends in great part on citations and references made possible via this data paper. One specific reuse of the data serves as an integral part in GreyNet's workshops and seminars on Data Papers^{5,6} providing comparative results, a tested module and template, as well as a use case that has the potential for implementation in other fields science and communities of practice. Nevertheless, as with long tail research data⁷ limitations may rest in its size and scale. Forty-four respondents do not actually allow for an accepted expression of results in percentages. Also, the original data remains on the server of SurveyMonkey for a limited time, while a copy is stored and preserved in the DANS EASY Archive. It is worth mentioning that this archive carries the CoreTrustSeal⁸, which likewise signals that the data complies with the FAIR Principles⁹.

To date, only the data and metadata are accessible in the DANS EASY Archive. However, the data specific to this survey as well as its comparison with data from the 2011 survey both stand to benefit from a forthcoming analysis. This will be deposited as a preprint under the same DOI and becomes a component in the enhanced publication.

Linked References

¹ Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication
<http://greyguiderep.isti.cnr.it/linkdoc.php?idcode=2018-GL20-01&authority=GreyGuide&collection=GLA&langver=en>

² Conference Proceedings on Grey Literature: GL Conference Series. ISSN 1386-2316 Print, ISSN 2211-7199 PDF
<http://greyguide.isti.cnr.it/gl-proceedings-2003/>

³ SurveyMonkey, A Survey Platform <https://www.surveymonkey.com/home/>

⁴ DANS EASY Archive <https://easy.dans.knaw.nl/ui/home>

⁵ The GreyForum Series - Where grey literature provides common ground for information professionals and practitioners in the process of knowledge transfer. <http://www.greynet.org/greyforumseries.html>

⁶ GreyWorks - Summer Workshop Series <http://www.greynet.org/traininglab/greyworks.html>

⁷ Long tail of research data <https://www.radar-projekt.org/display/RE/Glossar#Glossar-Longtailofresearchdata>

⁸ CoreTrustSeal - Core Trustworthy Data Repositories Requirements
<https://www.coretrustseal.org/wp-content/uploads/2018/04/DANS-Electronic-Archiving-SYstem-EASY-.pdf>

⁹ FAIR-Data Principles <https://www.force11.org/group/fairgroup/fairprinciples>

Data from:

“Leveraging Grey Literature: Capitalizing on Value and the Return on Investment”¹

<https://doi.org/10.17026/dans-zbf-kqwj>

Dominic Farace, GreyNet International, Netherlands

<https://orcid.org/0000-0003-2561-3631>

Project Administration/ Dataset Creator / Draft Text

Jerry Frantzen, GreyNet International, Netherlands

<https://orcid.org/0000-0002-3405-7078>

Visualization

Stefania Biagioni, Institute of Information Science and Technologies, Italy

<https://orcid.org/0000-0001-9518-0267>

Reviewer

Carlo Carlesi, Institute of Information Science and Technologies, Italy

<https://orcid.org/0000-0001-9808-6268>

Reviewer

Abstract

The data collected is based on the answers of 77 respondents to an online questionnaire carried out among the diverse population of stakeholders within the international Grey Literature Network Service, GreyNet.org². The dataset consists of ten questions, the last of which is open ended and all of which allow for additional comments. The first question asks the respondents to identify in which capacity or capacities they see themselves as stakeholders in GreyNet. The following eight questions allow them to indicate their level of use or participation in the organization’s sustained information resources. The data was collected and computer analyzed via SurveyMonkey, where it remains stored along with a copy in .ods format housed in the DANS Easy Archive. The data was collected in 2015 within a project on Leveraging Grey Literature. While the data was collected three years ago it offers a valuable baseline from which the organization can compare and reassess the degree to which its sustained resources are used by the stakeholders.

Keywords: Leveraging; Sustained Resources; Stakeholders; Use Case

Subject Area: Information Science

Methods Applied

● Steps

In producing the dataset an online questionnaire was drawn-up and implemented using the freeware ‘SurveyMonkey’³. The questionnaire contains ten queries each preceded by a brief statement pertaining to one of the sustained resources in the study. This was intended to allow for more informed responses. The final, open-ended question allows the opportunity to contact the respondents for a follow-up interview as well as provide them firsthand with the results of the survey.

● Sampling strategy

The population of the survey was opened for 12 days to recipients of GreyNet’s Distribution List, its LinkedIn members⁴, and Twitter followers⁵. The population of the survey was not numerically determined prior to its implementation. Hence, the total number of respondents to the questionnaire was used to calculate the results of each survey question except for the tenth question, which was open-ended. The results ranged from 77 (100%) to 62 (80.5%).

● Quality Control

While there was no specific control carried out on the data acquired from the survey, it should be noted that none of the categories within a given response were grouped or otherwise normalized. And, while only

29 (37.7%) of the 77 survey respondents chose to provide their contact details, no apparent cause is given to question the validity of the responses.

Dataset Description

File name:	Survey Results Leveraging.ods
Format name and version:	ODS 2018
Creation dates:	from 2015-06-19 to 2015-06-30
Language:	English
License:	CC0 Waiver - no rights reserved
Archive name:	DANS EASY Archive
Publication date:	2016-03-24

Potential Reuse of the Data

This data collected three years ago continues to provide a base of reference for reuse and further analysis. The seven types of stakeholders and nine sustained information resources selected for the survey in 2015 still remain the same. It is considered that the survey data would not only serve GreyNet.org in a follow-up study providing comparative results and enabling assessment, but it would also offer other communities of practice an applied method along with sample results. The method employed and the sample results have already been incorporated as a use case in two workshops^{6 7} on leveraging information resources.

On a more critical note, the data is limited by the number of respondents – 77 in total. This number does not allow for significant statistical results. In fact, the number itself does not formally allow for the expression of results in percentages. Also, the original data remains within the domain of SurveyMonkey, while only a copy is stored in the DANS EASY Archive⁸. However, it is worth mentioning that this archive carries the CoreTrustSeal⁹, which likewise signals that the data complies with the FAIR-Principles¹⁰.

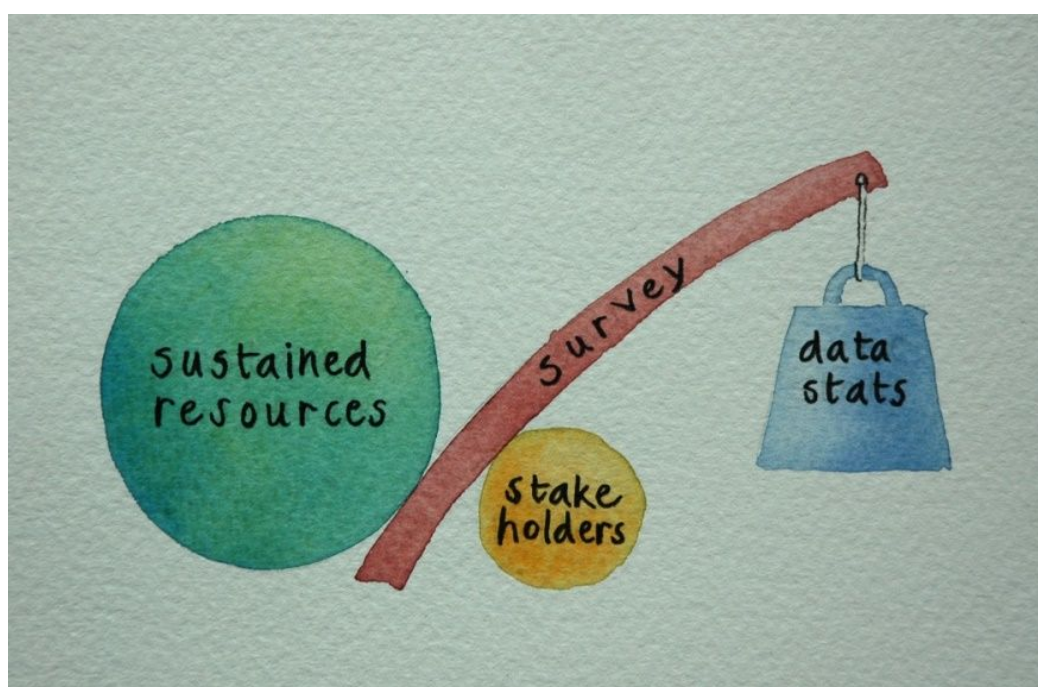


Illustration – Leveraging Sustained Resources

Linked References

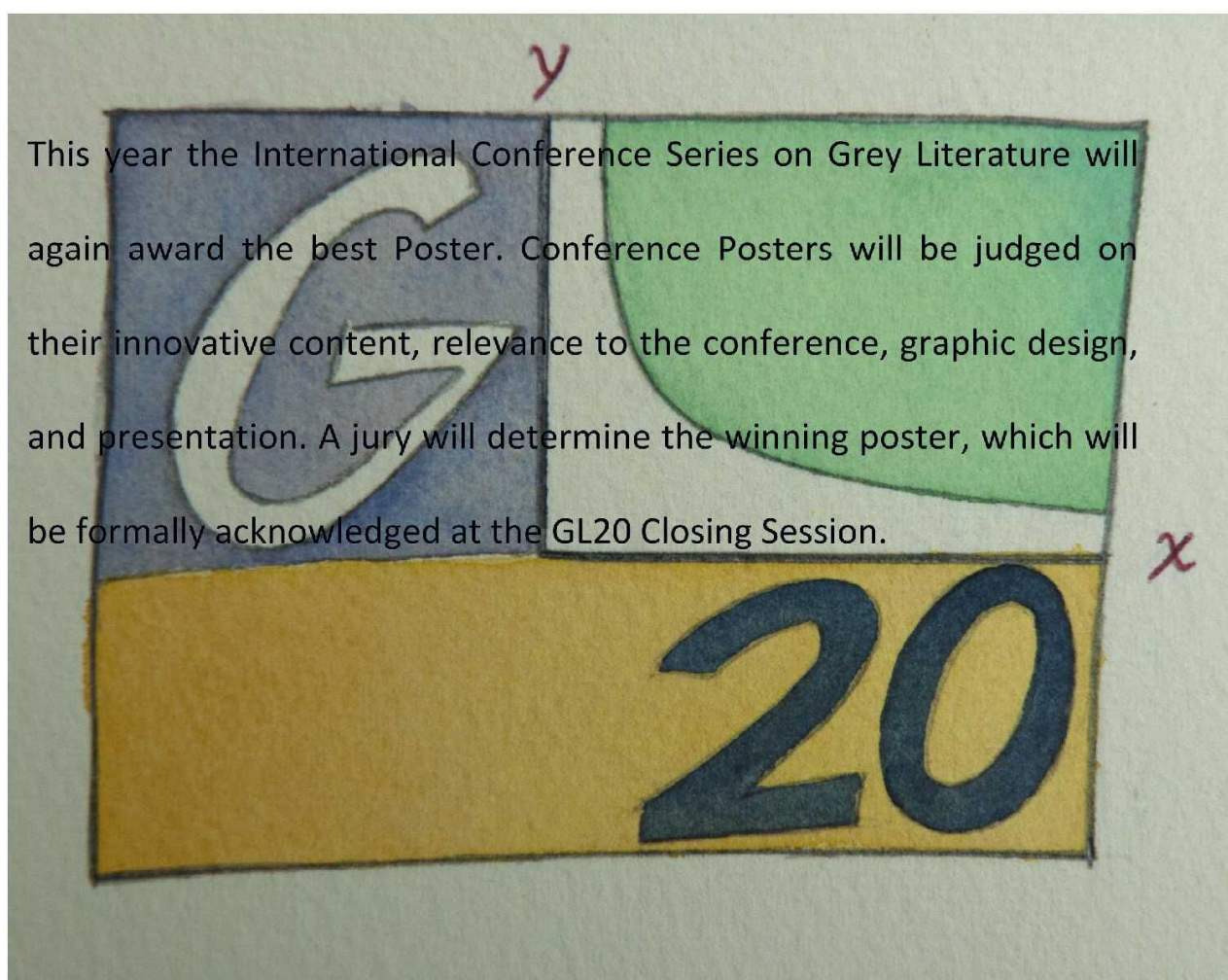
- ¹ Farace, D., [et al] 2016. Leveraging Grey Literature – Capitalizing on Value and the Return on Investment: A Cumulative Case Study. Seventeenth International Conference on Grey Literature.
<http://greyguiderep.isti.cnr.it/linkdoc.php?idcode=2016-GL17-020&authority=GreyGuide&collection=GLP&langver=en>
- ² GreyNet International <http://www.greynet.org/>
- ³ SurveyMonkey, A Survey Platform <https://www.surveymonkey.com/home/>
- ⁴ GreyNet LinkedIn <https://www.linkedin.com/groups/3718857>
- ⁵ GreyNet Twitter <https://twitter.com/GreyLitNet>
- ⁶ GreyWorks 2015 [www.greynet.org/images/GreyWorks_2015 - Leveraging Grey Literature.pdf](http://www.greynet.org/images/GreyWorks_2015_-_Leveraging_Grey_Literature.pdf)
- ⁷ GreyWorks 2016 http://www.greynet.org/images/GreyWorks_2016_Flyer_.pdf
- ⁸ DANS EASY Archive <https://easy.dans.knaw.nl/ui/home>
- ⁹ CoreTrustSeal <https://www.coretrustseal.org/wp-content/uploads/2018/04/DANS-Electronic-Archiving-SYstem-EASY-.pdf>
- ¹⁰ FAIR-Data Principles <https://www.force11.org/group/fairgroup/fairprinciples>

Twentieth International Conference on Grey Literature

Research Data Fuels and Sustains Grey Literature

Loyola University New Orleans, USA • December 3-4, 2018

BEST CONFERENCE POSTER 2018



Participants who seek to present a poster during GL20 are invited to submit an English abstract. The abstract should describe the project and/or related information product or service. The abstract should likewise include a title, name(s) of the presenter(s) and their full address information. Abstracts are an important source of information available prior to the conference accessible to both the conference delegates and participants. For more information, <http://www.textrelease.com/gl20callforposters.html>

TextRelease

GL20 Program and Conference Bureau

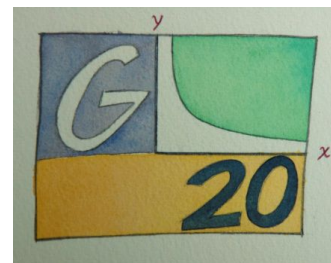
Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com • conference@textrelease.com
 Tel/Fax +31-20-331.2420

Twentieth International Conference on Grey Literature Research Data Fuels and Sustains Grey Literature



Loyola University, New Orleans, Louisiana, USA

December 3-4, 2018



Call for Posters

Title of Poster:	Conference Topic(s):
Presenter Name(s):	Organization(s):
Postal Address:	Phone:
Postal/Zip Code – City:	Email:
Country:	URL:

Participants who seek to present a poster during GL20 are invited to submit an English abstract between 250-300 words. The abstract should describe the project and/or related information product or service. The abstract should likewise include a title, name(s) of the presenter(s) and their full address information. Abstracts are an important source of information available prior to the conference that is accessible to both the delegates and conference participants.

Related Topics

- ☐ Long Tail Research Data and Grey Literature
- ☐ Metadata on the Frontline of Research
- ☐ Data Management and the Role of Librarians
- ☐ Current Research Trends in Grey Literature
- ☐ Research Data and Open Access Compliance
- ☐ Data Policies Reflect the Political Will of Organizations
- ☐ Other related topic:

Due Date, Format, and Poster Prize 2018

Timely registration is a guarantee of your placement on the conference program. The conference venue is able to accommodate a limited number of physical posters (20 inches x 30 inches / 50,8 cm x 76.2 cm), which will be mounted on foam board.

Abstracts in MSWord should be emailed to conference@textrelease.com on or before **October 30th 2018**. Those submitting poster abstracts will receive verification upon receipt and this will be accompanied by further guidelines for posters, submission of PowerPoint slides, Conference Registration, etc.

Those presenting conference posters are eligible for the Poster Prize 2018 that will be awarded during the GL20 Closing Session. Posters will be judged on their innovative content, relevance to the conference, graphic design, and presentation.

TextRelease

GL20 Program and Conference Bureau

Javstraat 194-HS, 1095 CP Amsterdam, The Netherlands

www.textrelease.com • conference@textrelease.com

Tel. +31-20-331.2420

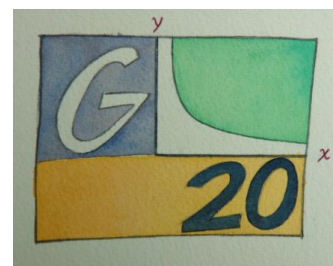
Twentieth International Conference on Grey Literature

Research Data Fuels and Sustains Grey Literature



Loyola University, New Orleans, Louisiana, USA

December 3-4, 2018



Conference Registration

The Conference Registration Fee includes attendance during the Plenary, Panel, and Poster Sessions, a copy of the GL20 Program Book, the full-text of the Conference Papers and PowerPoint Slides along with the conference pouch and participant badge. Also included are lunches and refreshments during the morning and afternoon breaks.

Payment after October 15th 2018 add a US\$ 50 Surcharge

☐ US\$ 495 Participant Fee ☐ US\$ 395 Author Fee ☐ US\$ 275 Day Pass ☐ US\$ 100 Student

☐ *GreyNet* Member entitled to a 20% reduction on the conference fee

☐ Affiliate of Loyola University is entitled to a 20% reduction on the conference fee

Registrant's Name:

Organization:

Address/P.O. Box:

Postal/Zip Code - City, Country:

Telephone:

Email:

Method of Payment:

☐ Direct bank transfer to TextRelease, Rabobank, Amsterdam, The Netherlands
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42 with reference to GL20 and Registrant's Name

☐ MasterCard ☐ Visa Card ☐ American Express

Card No. _____ Expiration Date: _____

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

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

Place _____ Date _____

NOTE Credit Card transactions will be authorized via Ogone/Ingenico, a designated payment service for Visa and MasterCard.

TextRelease

GL20 Program and Conference Bureau

Javstraat 194-HS, 1095 CP Amsterdam, The Netherlands

www.textrelease.com • conference@textrelease.com

Tel. +31-20-331.2420

Author Information

Balashova, Yuliya B.
119

Professor Yuliya B. Balashova, is Doctor of Philology, journalist, Associate Professor at St Petersburg State University, Russia. Email: j.balashova@spbu.ru

Biagioni, Stefania
151

Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in Data Processing and DBMS. She is currently a member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR) located in Pisa. She is head librarian of the Multidisciplinary Library of the CNR Campus in Pisa and member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been the responsible of ERCIM Technical Reference Digital Library (ETRDL) and currently of the PUMA (PUBlication MANagement) & MetaPub, a service oriented and user focused infrastructure for institutional and thematic Open Access repositories looking at the DRIVER/OpenAire vision, <http://puma.isti.cnr.it>. She has coauthored a number of publications dealing with digital libraries. Her activities include integration of grey literature into library collections and web access to the library's digital resources, including electronic journals and databases. She is a member of GreyNet since 2005. Since 2013 she is involved on the GreyGuide Project. Email: stefania.biagioni@isti.cnr.it

Cancedda, Flavia
143

Flavia Cancedda, Technologist, librarian and information specialist. Head of the Italian ISSN Centre (Central Library CNR); Component of "Documentation and Information" Committee of the UNI, National standardization organism (Chairman of the Subcommittee "Presentation, identification and description of documents"; component of the Working Group on National Standard about Professional Requirements for Librarians); Component of the ACNP Libraries Committee (Archivio Collettivo Nazionale Periodici) and of the ACNP Cataloguing Rules Revision Working Group. Since 1991 to 2000 librarian in two different Italian universities and in the National Library of Florence (section: National Bibliography); since 2001 librarian in the Central Library of the Italian National Research Council (CNR). Master's degree in Libraries Management and Direction. Post-graduate degree as Manuscripts conservator at the Special School for Archivists and Librarians (La Sapienza University in

Rome). Author of books and essays about the history of libraries and historical bibliography.

Email: flavia.cancedda@cnr.it

Carlesi, Carlo
151

Carlo Carlesi graduated in Computer Science, worked since 1970 at the IEI (now ISTI) of the Italian National Research Council in Pisa. His interest are focused in many software engineering field such as: Development of data base systems, Software quality and testing, Administration and UNIX system management, Digital library systems, Network security and management. In the year 2000-2010 he was head of the "Information Technology Security Service" of the Institute. He participated in many national and international projects, the last being: Italian Project for Research in the Antarctic - (South Pole project), the aim of the project was the creation of a Multidisciplinary Integrated Information System to manage and query the Set of Antarctic Data Bases; ERCIM Technical Reference Digital Library - (ETRDL Project), The Digital Library service allows public access through Internet to the technical reports and other not published document produced by several organizations. It is currently a Research Associate of the Institute ISTI and he is involved in the following projects: PUMA - Publication Management. The Digital Library service allows public access (when permitted) through Internet to the published documents produced by CNR Organizations. GreyGuide - Guide to Good Practices and Resources in Grey Literature. An online forum and repository of good practice and resources in the field of grey literature. Email: carlo.carlesi@isti.cnr.it

De Biagi, Luisa
143

Luisa De Biagi got her Laurea Degree in Literature and Philosophy at 'La Sapienza' Univ. of Rome (Art history and Cultural heritage). With a Specialization in 'Archivist-Palaeographer' (Vatican School of Palaeography, Diplomatics and Archivistics at the Vatican Secret Archive) as well as a Specialization Degree in Archivistics, Palaeography and Diplomatics (Archivio di Stato, Rome) and a Degree from the Vatican School of Library Sciences. De Biagi further holds a Master in 'Business Publishing' (LUISS Management – Rome). She's been working for the SIGLE Network (System for Information on Grey Literature in Europe) since 2002, Since 2010 she's is responsible for the Italian National Referring Centre of Grey Literature at CNR Central Library 'G. Marconi' as representative of the European Network and

Author Information CONTINUED

Openarchive OpenGrey (System for Information on Grey Literature in Europe). She's taken part in 3 editions of the Annual International Congress on Grey Literature – GL (GL5, Amsterdam, GL13, Washington d.c. GL14, Rome and GL15 at Bratislava). She's is also a member of the CNR Working Group for Cedefop-Refernet Project (Consortium for Professional Education and Training coordinated by ISFOL), the Committee for Legal Deposit Acquisition at CNR Central Library, and a member of the European Association of Health Information and Libraries (EAHIL). She's also responsible for the Library Functional Units 'Education and Training' and 'Cultural Activities Management', organizing didactics laboratories for students, professional training courses and teaching in professional trainings for librarians, students and users.

Email: luisa.debiagi@cnr.it

Farace, Dominic **149, 151**

Dominic Farace is Head of GreyNet International and Director of TextRelease, an independent information bureau specializing in grey literature and networked information. He holds degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht, The Netherlands, where he has lived and worked since 1976. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences (SWIDOC/KNAW), Farace founded GreyNet, Grey Literature Network Service in 1992. He has since been responsible for the International Conference Series on Grey Literature (1993-2013). In this capacity, he also serves as Program and Conference Director as well as managing editor of the Conference Proceedings. He is editor of The Grey Journal and provides workshops and training in the field of grey literature.

Email: info@greynet.org

Frantzen, Jerry **151**

Jerry Frantzen graduated in 1999 from the Amsterdam University of Applied Sciences (HvA) in Library and Information Science. Frantzen is the technical editor of The Grey Journal (TGJ). And, since 1996, he is affiliated with GreyNet, Grey Literature Network Service, as a freelance technical consultant. Email: info@greynet.org

Pavlov, Leonid P. **123**

Graduated from Moscow Physical-Engineering Institute (now - National Nuclear Research University), Dipl. Eng. in computer systems. PhD in informatics. Main field of scientific interest - grey literature, information systems. Since 1976 has been with the Scientific and Technical Information Centre of Russia – the federal GL centre, in 2009 merged with the Centre of Information Technologies and Systems for Executive Power Authorities (CIT&S). Participated in the R&D, implementation and management of nation-wide information systems for grey literature as scientific researcher, project manager, acting director, specialist-in-chief. Author of more than 110 scholarly works (including those written in English and published abroad).

Email: pavlov@vntic.org.ru

Prost, Hélène **113**

Hélène Prost is information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiICO research laboratory (University of Lille 3). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, document delivery, usage analysis, grey literature and open access, and she is author of several publications.

Email: helene.prost@inist.fr

Schöpfel, Joachim **113, 149**

Joachim Schöpfel is senior lecturer at the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiICO laboratory. He is interested in scientific information, academic publishing, open access, grey literature and eScience. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. Email: joachim.schopfel@univ-lille3.fr

Notes for Contributors

Non-Exclusive Rights Agreement

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

Submission Requirements

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

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

REFERENCE GUIDELINES

General

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

Specific

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

Examples

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

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

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

Review Process

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

Journal Publication and Article Deposit

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

An International Journal on Grey Literature **'PROFILING GREY DATA FOR PUBLIC ACCESS'**

D4Humanities: Deposit of Dissertation Data in Social Sciences & Humanities – A Project in Digital Humanities	113
<i>Joachim Schöpfel and Hélène Prost (France)</i>	
Public Access to the Dissertations in Russia	119
<i>Yuliya B. Balashova (Russian Federation)</i>	
Data on Grey Literature Containing Scientific and Technical Information	123
<i>Leonid P. Pavlov (Russian Federation)</i>	
Indexing grey multilingual literature in General Practice: Family Medicine in the Era of Semantic Web	125
<i>Marc Jamoulle [et al] (Belgium)</i>	
Apps & Codes: Making profiles for fluid publishing contents	143
<i>Flavia Cancedda and Luisa De Biagi (Italy)</i>	
Data from “Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication”	149
<i>Dominic Farace (Netherlands) and Joachim Schöpfel (France)</i>	
Data from “Leveraging Grey Literature: Capitalizing on Value and the Return on Investment”	151
<i>Dominic Farace and Jerry Frantzen (Netherlands); Stefania Biagioni and Carlo Carlesi (Italy)</i>	

Colophon	110
Editor's Note	112
On The Newsfront	
Best Conference Poster 2018	154
GL20 Call for Posters - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	155
GL20 Registration - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	156
Advertisements	
KISTI, Korea Institute of Science and Technology Information	122
EBSCO LISTA - Library, Information Science & Technology Abstracts with Full Text	142
Author Information	157
Notes for Contributors	159