8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

CHALLENGES IN PROVIDING UNPUBLISHED RESEARCH DATA IN BIOMEDICINE TO GREY LITERATURE REPOSITORIES

Pavla Francová[1], Stephanie Krueger[1,2]

pavla.francova@techlib.cz, stephanie.krueger@techlib.cz

[1] National Library of Technology, Prague, Czech Republic

[2] University of Chemistry and Technology, Prague, Czech Republic; Humboldt-Universität zu Berlin

Abstract

Regardless of the scientific field or focus, every researcher produces a multitude of unpublished research data during his or her career, including project diaries, project proposals, laboratory notes, and also the outputs of collaboration between researchers – for example, images, video, and measurements. Such "ephemeral" data can be crucial in inspiring new research directions and perspectives, and is often currently not shared in open repositories, although making such data more accessible undoubtedly has value for researchers. While not "literature" according to the traditional definition, such contextual materials and data – at least in bioengineering and biophysics – are the de facto bases for any grey literature produced in such fields and are directly relevant when discussing the utility of a grey literature repository in relation to such research endeavors. In this paper, the authors describe the difficulties posed in searching for grey literature and data on a specific bioengineering topic, Magnetic

Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast method, and also provide examples of unindexed grey literature and data produced by scholars in this field.

Keywords
Grey Literature, Data Repository, Dark Data, Research Data, Magnetic Resonance Imaging, Magnetization Transfer Contrast, Biomedical Engineering, Biomedicine, Lung

---

Introduction
In this paper, the authors describe the difficulties posed in searching for grey literature on a specific bioengineering topic, Magnetic Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast (MTC) method. They also provide examples of unindexed grey literature and data produced by scholars in this field. Via this method, the authors combine (in case study format) the perspectives of information scientist and active researcher in order to address the broader questions about the current status of the accessibility of scholarly outputs in bioengineering based on a real-life example of information use and retrieval in this field, together with a description of scholarly outputs invisible to the outside world prior to the creation of what might be considered grey literature (e.g., conference paper or pre-print).

The authors identify a lack of accessibility to grey literature and data in this specific field which pose challenges for conducting research into this specific bioengineering topic and provide a list of possible areas in which libraries and/or scientists themselves might enhance accessibility to research outputs in the future.

The Bioengineer's Perspective: Searching for Literature and Data (including Grey Materials)
In this section, the authors describe a "real-life" search for literature and data as an example of information retrieval from the perspective of a researcher in the field of bioengineering utilizing the topic "Magnetic Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast method." The description of the development of a search query and the information retrieved represents an actual research session conducted by one of the contributing authors in October 2015 during a recent extension of her research project which was accepted at the University of Würzburg's Pulmonary Imaging Network (PINET) research group under European Union grant FP7-ITN and Marie Skłodowska-Curie Actions (MSCA) grant PITN-GA-2010-264864. The description of this session provides a starting point for future research into the accessibility of grey literature for scholars conducting research in this topical area. It is intended to address, by means of a real example, the paucity of understanding of the information-related behaviors in the sciences by describing, from the emic perspective of an actual working scientist, the steps involved in conducting research in her field. [A]

The following information retrieval steps (A-D) represent, according to this bioengineer, the way in which she typically conducts a search for literature and data, in which she begins her search using published scholarly outputs and – only in Steps C and D – opens up her search to include the outputs of pre-prints, grey literature, and dark data [B], the so-called "long tail of science" [C, D].

- **Pre-Step**: Develop search query
- **Step A**: Search fully-indexed bibliographic databases for relevant information and data
- **Step B**: Search for full-text in fully-indexed databases for relevant information and data
- **Step C**: Search institutional repositories (fully- or partially-indexed; this varies) for relevant information and data
- **Step D**: Search other grey and dark data repositories and resources.

In the following sections, the authors describe how the researcher typically creates a query set and the results of applying query variants to different kinds of data sources (i.e., steps A-D). For each step, the authors then comment on the availability of grey literature and data, from the perspective of the researcher.

Developing a Search Query Set

In medicine and biomedical engineering, researchers in this researcher's field are – in her opinion – not usually familiar with the MeSH (Medical Subject Headings) thesaurus, a controlled vocabulary for indexing articles utilized by PubMed (1[1]) (the free search engine for accessing the MEDLINE database provided by the United States National Library of Medicine and the US National Institutes of Health). However, the authors decided in this case study to use the MeSH controlled vocabulary to make their search queries more precise.

For this example, the researcher mapped her initial keyword set (Lung structure, Magnetic Resonance Imaging, and Magnetization Transfer Contrast Imaging) to the MeSH subject headings: "**Magnetic Resonance Imaging**", "**Lung,**" and "**Magnetization Transfer Contrast Imaging**" as well as several variants of the latter. The authors then utilized these subject headings for subsequent searches in different data sources. The variants of "**Magnetization Transfer Contrast Imaging**" allowed the authors to analyze the effect of phrase variants on the number of results for each resource.

[1] Available from **http://www.ncbi.nlm.nih.gov/pubmed**

A	"**Magnetic Resonance Imaging**" AND "**Lung**"
B	"**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**"
C	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**"
V1	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**"
V2	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**" OR "**Magnetization Transfer Imaging**"
V3	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**" OR "**Magnetization Transfer Imaging**" OR "**Magnetization Transfer Contrast Imaging**"

Table 1 Search terms developed using MeSH for the topic "Magnetic Resonance Imaging", "Lung," and several variations of "Magnetization Transfer Contrast"

Step A: Bibliographic Databases

Using these keywords, the authors retrieved results in the following bibliographic databases which are licensed by the National Library of Technology for its registered patrons (as is the case for the full-text databases mentioned in Step B below).

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
PubMed	8 275	2 213	17 [1-3]	225	617	617
SCOPUS	25 231	3 002	12 [1-4]	12	12	12
Web of Science - title	351	1 398	1 [3]	147	395	395
Web of Science - topic	3 293	3 462	10 [1-3,5]	458	1 161	1 161

Table 2 Number of results for queries A-V3 from Table 1 in key bibliographic databases. Phrase C provided the most relevant found articles. Search conducted October 2015.

Phrase C provided the most relevant articles according to the researcher, though the authors deemed only five as being directly relevant to the research topic. Unfortunately, no other kind of grey literature except the conference papers was included in these sources.

[1] ARNOLD, J.F.T. , M. KOTAS, R.W. PYZALSKI, E. D. PRACHT, M. FLENTJE, and P. M. JAKOB. Potential of magnetization transfer MRI for target volume definition in patients with non-small-cell lung cancer. *Journal of Magnetic Resonance Imaging* [online]. 2008, 28(6): 1417-1424 [cit. 2015-10-19]. DOI: 10.1002/jmri.21436.

[2] JAKOB, P.M., T. WANG, G. SCHULTZ, H. HEBESTREIT, A. HEBESTREIT, M. ELFEBER, D. HAHN, and A. HAASE. Magnetization transfer short inversion time inversion recovery enhanced1H MRI of the human lung. *Magma: Magnetic Resonance Materials in Physics, Biology, and Medicine* [online]. 2002, 15(1-3): 10-17 [cit. 2015-10-19]. DOI: 10.1007/bf02693839.

[3] KUZO, R.S., M.J. KORMANO, and M.J. LIPTON. Magnetization Transfer Magnetic Resonance Imaging of Parenchymal Lung Disease. *Investigative Radiology* [online]. 1995, 30(2): 118-122 [cit. 2015-10-19]. DOI: 10.1097/00004424-199502000-00011.

[4] NIEMI, P.T., M.E.S. KOMU, and S.K. KOSKINEN. Tissue specificity of low-field-strength magnetization transfer contrast imaging. *Journal of Magnetic Resonance Imaging* [online]. 1992, 2(2): 197-201 [cit. 2015-10-19]. DOI: 10.1002/jmri.1880020213.

[5] ARNOLD, J.F., M. KOTAS, D. PRACHT, M. FLENTJE, and P.M. JAKOB. Could Functional MRI Improve Radiation Therapy Planning in Non-Small Cell Lung Cancer? *International Journal of Radiation Oncology*Biology*Physics* [online]. 2005, 63: S224-S225 [cit. 2015-10-19]. DOI: 10.1016/j.ijrobp.2005.07.384.

Step B: Full-Text Databases

The authors then conducted a search across full-text subscription databases (see Table 3) utilizing the queries defined in Table 1.

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
EBSCOhost	2 235	834	4	44	208	67 371
ScienceDirect	42 465	5 860	300	300	300	300
SpringerLink **Biomedical Sciences**	5 872	815	83	156	322	322
SpringerLink Medicine	26 778	1 989	387	681	1 016	1 016
SpringerLink Public Health	1 215	138	45	65	76	76
Wiley Online Library	26 109	5 796	711	1 489	2 073	2 073
ProQuest Dissertations & Theses	9 483	6 681	282	855	2 317	2 317
ProQuest Health and Medicine	43 772	6 681	282	855	2 317	2 317

Table 3 Number of results for the query defined in Table 1 in selected full-text databases

Here, relevant results rarely included grey literature (i.e., project or technical protocols, or conference materials) or data. In the researcher's opinion, SpringerLink provided the most useful and relevant information in terms of grey literature because it indexes conference materials from field-related events organized by the European MRI society (ESMRMB), which enables access to reports about ongoing research projects [F] as well as conference abstracts and posters. In contrast, conference materials from another important scholarly society in this field, the International MRI society (ISMRM), are not indexed in these databases and therefore are only accessible to ISMRM members via login at the ISMRM websites and are not (as of October 2015) available to researchers at large who are not ISMRM members.

Step C: Institutional Repositories

The authors then conducted a search across selected institutional repositories, those with renowned research groups in this area of research (Table 4 below). Relevant grey literature results include conference materials (mostly posters and abstracts), dissertations, and occasionally project summaries or reports. Links to all repositories in Table 4 are available at The Directory of Open Access Repositories OpenDOAR (2).

(2) Available from **http://www.opendoar.org/countrylist.php**

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
Universität Würzburg	143	18	14	143	143	143
Friedrich-Alexander-Universität Erlangen-Nürnberg	88	7	3	88	88	88
Eberhard-Karls-Universität Tübingen	643	509	710	755	755	755
Forschungszentrums Jülich	4	3	0	0	0	0
Ruprecht-Karls-Universität, Heidelberg	161	2	23	0	0	0
Health Services Research Projects in Progress	8	0	0	0	0	0

Table 4 Number of results in selected institutional repositories using query defined in Table 1

Step D: Dedicated Grey Literature Repositories

Finally, the authors conducted a search across dedicated grey literature repositories. Most useful for this particular area of inquiry, from the researcher's perspective, are the **Public Health Grey Literature Sources (3)** (listed in Table 5) provided by the OPHLA (Ontario Public Libraries Association), which covers American, Canadian, and international grey data repositories; and the **Data Sharing Repositories** (4) provided by US National Library of Medicine.

International European Repositories	A	B	C
Electronic Theses Online Service (ETHOS) | British Library	22	2	0
Center for Research Libraries Foreign Dissertation	537	1	538
DART-Europe E-theses Portal	30	18	30
National Institute for Health and Clinical Excellence (NICE)	24	0	0
Public Health England	1	0	0
UK Department of Health	22	172	95
Nature Precedings	15	1	0
World Health Organization	93	0	0

Table 5 Number of results for query from Table 1 in dedicated grey literature repositories.

(3) Available from **http://www.ophla.ca/pdf/Public%20Health%20Grey%20Literature%20Sources.pdf**
(4) Available from **https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html**

Because of the narrow nature of this research query, the authors found no relevant results in GreyNet International, European Commission (EUROPA) Public Health or European Union Open Data Portal.

In sum, the grey literature and data results from all four sample searches yielded little relevant information and results were primarily conference proceedings and related materials. Raw data of other ephemeral materials such as laboratory notes relevant to this project was not available from any resource.

In the following section, the authors describe the particular difficulties faced by researchers in this field regarding data and related ephemeral materials, which is – as seen in the examples above – currently quite inaccessible to scholars in traditional databases as well as in repositories.

The Torturous Path from Dark to Grey Data

Any research conducted in this particular area of bioengineering research results in various data outputs. For this researcher, data she commonly works with for one project can be divided into the following groups (Table 6), with related storage requirements indicated in the far-right column.

Data Type	**Size**
Single data set (i.e., an individual MRI image received using a measuring protocol)	**3 - 4 MB**
RAW data sets (total images per scientific project)	**80 - 100 GB**
Laboratory notes and diaries (evaluation of single data sets, parameters)	**MB**
Summaries and statistics (comparison of data sets per chosen parameter)	**MB**
Conference materials (posters, presentation, supportive materials)	**2 - 3 GB**
Supportive materials for peer-reviewed outcomes (images, tables, graphs)	**MB**
Programming files (measurement and evaluation files, necessary .exe programs)	**35 - 60 GB**
Research-related literature and data (including full-text results of literature search, text variants, images etc)	**2 GB**
Total size of all the related project materials	**100 - 170 GB**

Table 6 List of different scientific data outputs and their file sizes (based on the realized scientific project); in several cases, only list as megabits (MB) because it is impossible to calculate their exact size in comparison with other data types within the context of this paper.

Other valuable and even essential data sources include records of development for the project's working hypotheses, false measurements (human versus technical errors), and RAW data with unusual artifacts.

If these kinds of data (and similar data for related projects) were more broadly available to researchers in this field, direct benefits for researchers would include: detailed information about the chronology and methodology of a particular project, how the project was evaluated,

possible pitfalls future researchers might avoid, and areas in which future research is not yet possible because of "dead ends" (e.g., inability of current technologies to address particular research questions). In particular, a deeper understanding of the challenges of a chosen method and verification of hypotheses against previous experiments might be particularly useful if such data were more readily available.

Further benefits of accessible RAW data images would include, for MRI programmers, the ability to test data sets in order to improve evaluation software (e.g., to compare MRI images against previous results). Also, researchers would be able to additionally compare results from other students according to selected parameters even if the original author did not evaluate them (e.g., influence of sex or age, ventilated gas, breath or cardiac phase, etc.). This detailed data would enable more thorough and reliable statistics regarding parameters, data, and even artifacts.

Laboratory diaries and notes offer a unique complex perspective into each particular scientific project. Especially when conducting basic science, even small notes about what was manageable or what failed can spare other researchers tremendous amounts of time. However, many small side experiments might have a great value and can contribute to deeper understanding topics, yet they are currently mostly stored only in researchers' heads and officially are not made public.

Of course, the strong experiment-dependence of the Magnetization Transfer Contrast methods might make comparison with the results of different experiments impossible. But with detailed knowledge of ALL the measurement parameters which might be possible to observe using MTC techniques, similarities and regularities not obvious at first (or even third) glance might be actually comparable for a carefully-chosen parameter.

Many of the challenges of publishing such dark data, on the other hand, are not so obvious. When you have terabits of dark data, where might one (publicly and accessibly) store them? And how should one properly index them to make them accessible? And – particularly relevant in biomedicine and biomedical engineering – how might one address the ethical questions regarding human (i.e., volunteer or patient) data? No potential data repository in this narrow field currently provides satisfactory answers to these questions.

Another complication is the manner in which one might present individual data set results, laboratory notes, etc. In this field, there are not currently any standardized formats or platforms which might broaden access to these materials. Because of this, every researcher currently must find a compromise between the presentation of original data (e.g., a written lab diary) and its public presentation (experiment's records or technical protocols). This extra work, combined with a need for highly structured texts and outcomes, discourage many scientists from sharing such data. There is currently no universal platform or repository for sharing such information in this field.

Illustrations of Typically Unindexed Information (DARK DATA)

As a brief introduction for non-scientists, the authors would like to provide in this section examples of real project outcomes: MRI images (full single data set, images with artifacts; Fig. 1 and 2) and sample scientific laboratory notes (Fig. 3). Fig. 4 provides an overview of the data summary.

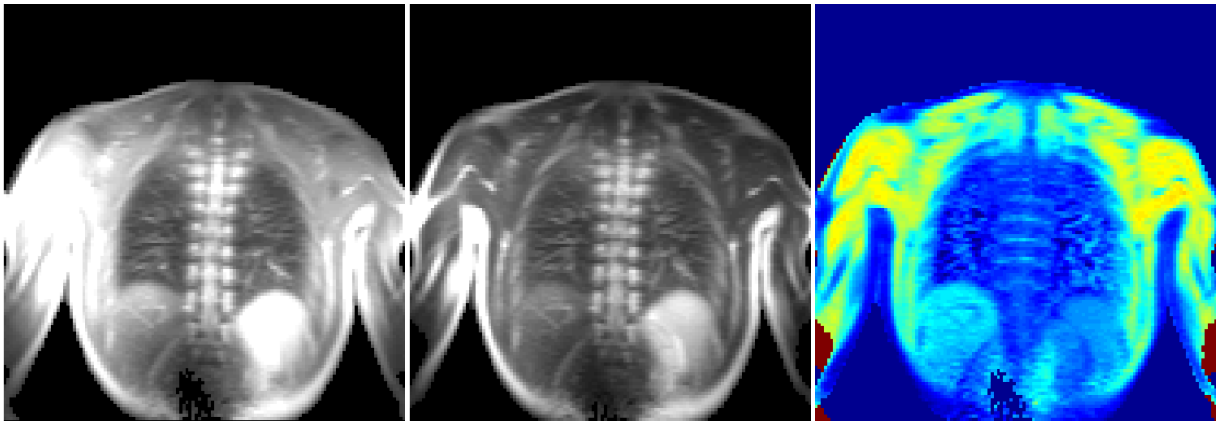

Figure 1 Typical example of MTC-MRI evaluated single data set - MRI images without and with MTC preparation and the final calculated (colorful) contrast image (useful data set).

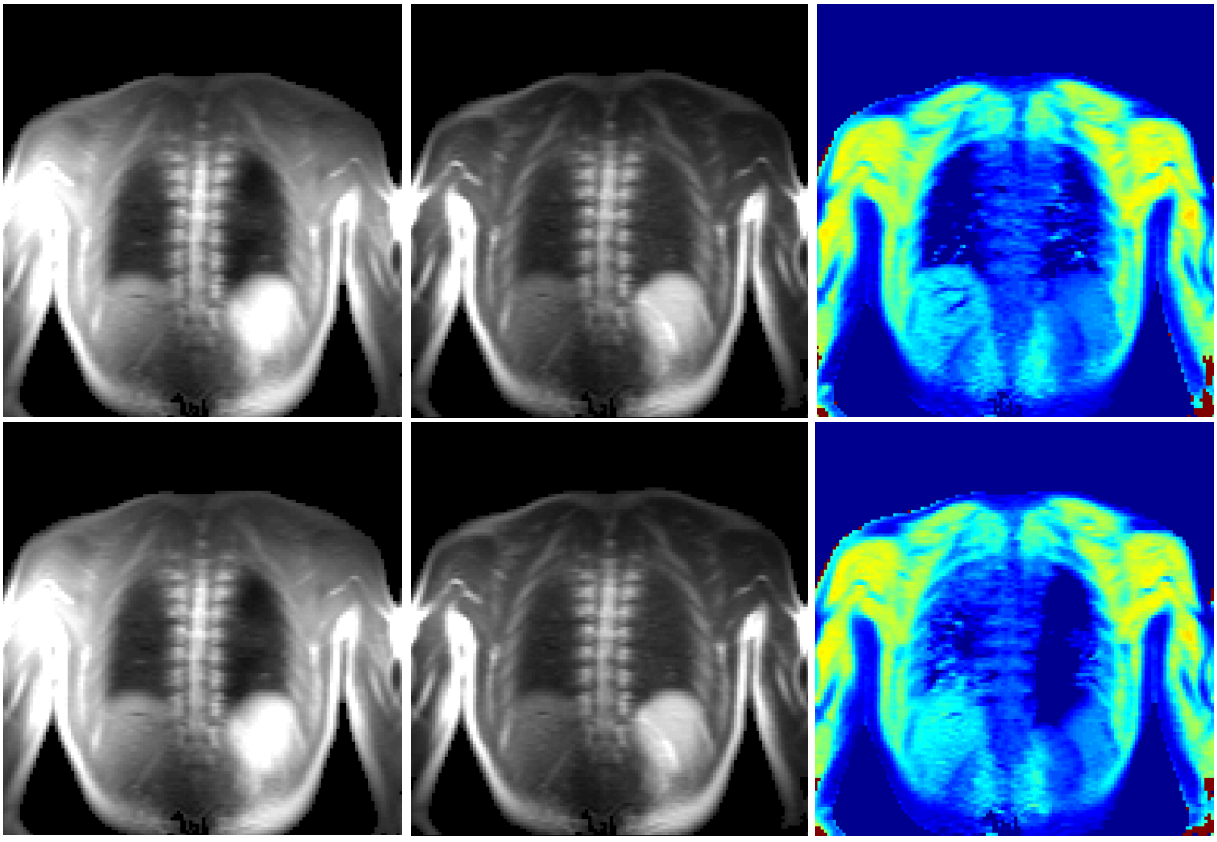

Figure 2 Two of typical examples of a severe artefacts in the final calculated MRI image (useless data set) - first set according to comparison from two images with different cardiac phase, second data set due to motion artefact.

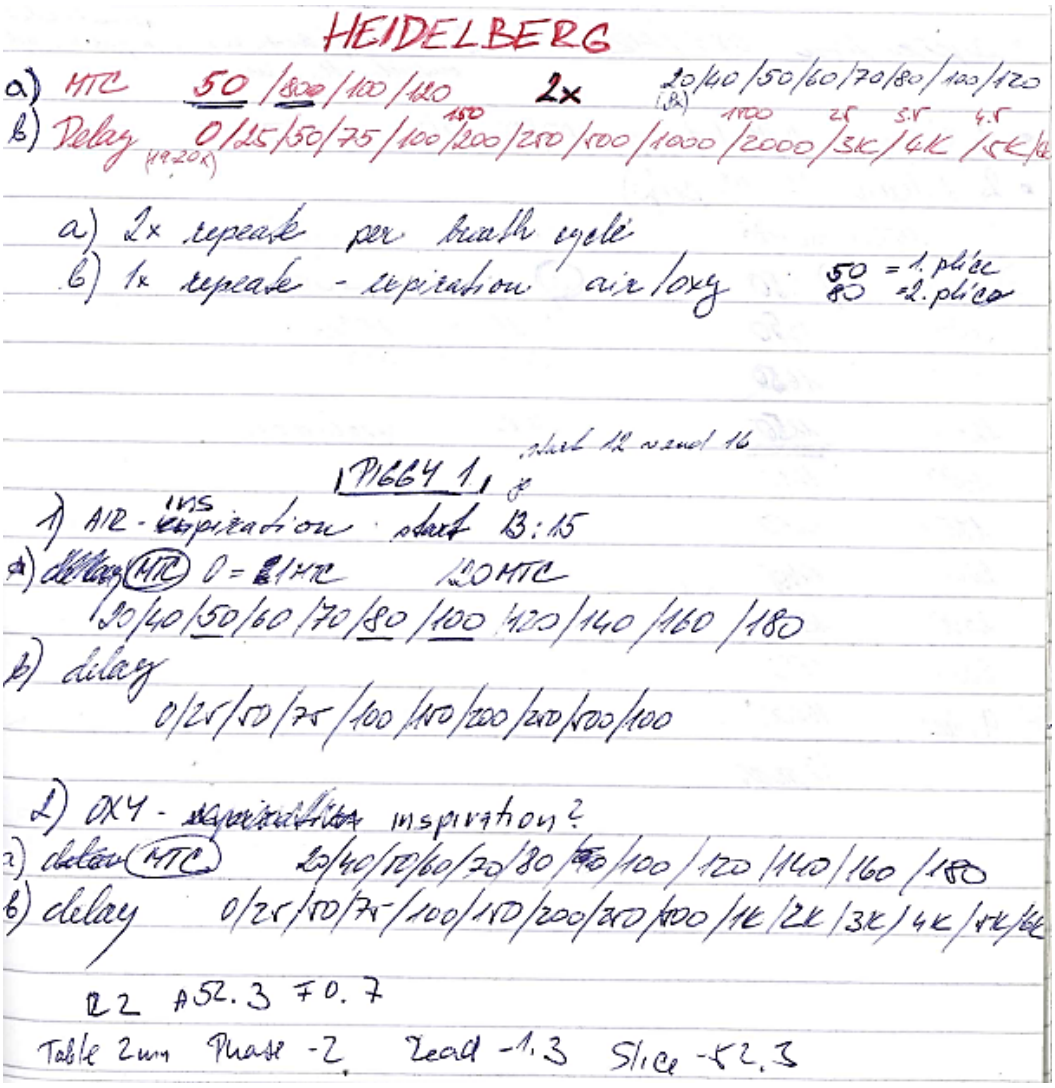

Figure 3 Illustrational scan of the laboratory diary with notes (hand-written, in current state unpublishable)

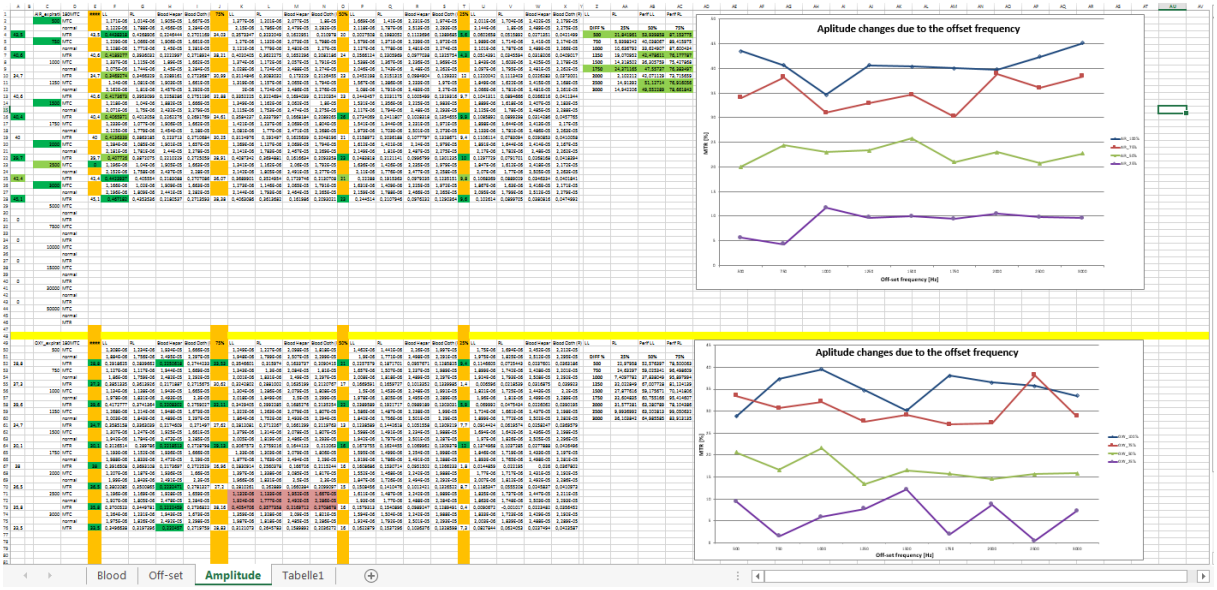

Figure 4 Illustrational scan of evaluated and roughly visualized individual data set's results (no personal data included) - shown results represent the measurement on the lung phantom with ex-vivo porcine lungs, ventilated with room air and pure oxygen.

Conclusion and Recommendations

In this case study, the authors provided examples of the availability and accessibility of grey literature and data in a new, very narrow research topic in medical diagnostics/bioengineering modeled on the real-life, completed project "Magnetization Resonance Imaging of the Lung Structure using Magnetization Transfer Contrast Method." All images and examples used belong to the authors and have not yet been previously published.

Grey literature identified in four sample searches across different resource types include conference materials (abstracts, posters and e-posters, presentations) and dissertations, but rarely project reports. For national institutional and other grey data repositories, language barriers potentially create great challenges; this paper focused on search results using English-language queries only and additional language queries would need to be tested in future studies, as they were beyond the scope of this paper.

In terms of grey literature and grey/dark data, while materials from conferences or scientific meetings as well as dissertations and projects reports from universities are available and can be found, other data types such as laboratory diaries, measurement reports, and case studies are – in this emerging field – completely inaccessible. Due to aforementioned challenges such as the overall storage requirements for dark data (especially RAW data), no standardized indexing formats, ethical guidelines, and lack of universal storage platforms in this field, only a few researchers are willing to share their dark data in grey literature repositories. The authors did identify a few bioinformatics pioneers who are interested in opening the doors of access to bioengineering data in the future and who envision grey data repositories which might include more than conference materials. Examples of recent projects in this area include XTENS [G], BIRN (4), and OpenScienceLink (OSL, (5)) [I]; the open source ***Biomedical Data Journal*** provides a forum for recent research in bioinformatics.

To improve the accessibility of grey literature and grey/dark data in this field, the authors recommend the following: first, it is necessary to determine which data sources could/should be stored in grey data repositories and prepare a universal format/platform in order to properly structure and describe data for future research. Second, ethical questions regarding the handling of personal information regarding human subjects (i.e., volunteers and patients) including preventing misuse or appearance to the public (non-medical) audience must be addressed. Until these basic questions are defined, it is unlikely that researchers in this particular field would be willing to share dark data more broadly.

(4) Available from **http://www.birncommunity.org/about/birn-video-intro/**
(5) Available from **https://www.gopubmed.org/web/oslplatform/**

References

[A] KRUEGER, S. *Beyond the Paywall: A Multi-Sited Ethnographic Examination of the Information-Related Behaviors of Six Scientists*. Berlin: Humboldt-Universität zu Berlin. Forthcoming dissertation to be published 2016.

[B] YOUNG, J. M. *An epidemiology of big data*. Syracuse: Syracuse University, 2014. Dissertation. ISBN 9781303909979. Available from ProQuest Dissertations & Theses Full Text: The Sciences and Engineering Collection.

[C] HEIDORN, P.B. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*. 2008, vol. 57, no. 2. pp. 280-299. ISSN 00242594. Available from ProQuest SciTech Collection.

[D] LANDER, H. and A. RAJASEKAR. *DataBridge: Creating Bridges to find Dark Data* [online]. Chapel Hill: RENCI (University of North Carolina), 2015 [cit. 2015-10-18]. RENCI White papers series, Vol. 3, No.5. DOI: 10.7921/G0MS3QNF.

[E] JIROTKA, M., C.P. LEE, and G.M. OLSON. Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW)* [online]. 2013, **22**(4-6): 667-715 [cit. 2015-10-18]. DOI: 10.1007/s10606-012-9184-0.

[F] *ESMRMB 2015: 32nd Annual Scientific Meeting, Edinburgh, UK, 1-3 October: Abstracts, Thursday* [online]. Springer Berlin Heidelberg, 2015 [cit. 2015-10-18]. ISSN 1352-5243. DOI: 10.1007/s10334-015-0487-2.

[G] IZZO, M., G. ARNULFO, M.C. PIASTRA, V. TEDONE, L. VARESIO, and M. M. FATO. XTENS - A JSON-based digital repository for biomedical data management (2015). In: F. ORTUNO and I. ROJAS, ed. *Bioinformatics and Biomedical Engineering: Third International Conference, IWBBIO 2015, Granada, Spain, April 15-17, 2015. Proceedings, Part II.* Springer International Publishing, 2015, p. 123-130. Lecture Notes in Computer Science, vol. 9044. ISBN 978-3-319-16479-3. DOI 10.1007/978-3-319-16480-9

[H] EISINGER, D., G. TSATSARONIS, A. PETROVA, E. KARANASTASIS, V. ANDRONIKOU, and E. CHONDROGIANNIS. OSL Platform: A Link to Open-access Scientific Information and Structured Data. *Biomedical Data Journal* [online]. 2015, **01**(1): 52-54 [cit. 2015-10-18]. DOI: 10.11610/bmdj.01109.