When is ‘grey’ too ‘grey’? - A case of grey data

Dr. Dobrica Savić
Nuclear Information Section, IAEA

Abstract

Conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security have become important concerns. Trustworthiness of news and information, and of grey and other literature types has become of interest to the public, as well as to many information science and technology researchers. Starting with a definition of grey literature, and continuing with white, dark and grey data, this paper concentrates mainly on grey data as an emerging grey literature data type and its various ‘shades’ of trust. Special attention is given to data in the context of grey systems theory, anonymous data, and unstructured and unmanaged data. Based on a review of relevant literature and current practices, trustworthiness of grey data is analysed and elaborated. Guidelines and warning signs of grey data trustworthiness are identified, and conclusions offered.

Keywords: grey literature, grey data

Dr. Dobrica Savić is Head of the Nuclear Information Section (NIS) of the IAEA. He holds a PhD degree from Middlesex University in London, an MPhil degree in Library and Information Science from Loughborough University, UK, an MA in International Relations from the University of Belgrade, Serbia, as well as a Graduate Diploma in Public Administration, Concordia University, Montreal, Canada. He has extensive experience in the management and operations of web, library, information and knowledge management, as well as records management and archives services across various United Nations Agencies, including UNV, UNESCO, World Bank, ICAO, and the IAEA. His main interests are digital transformation, creativity, innovation and use of information technology in library and information services.

Contact: *www.linkedin.com/in/dobricasavic*
ORCID ID: *orcid.org/0000-0003-1123-9693*

Why are we concerned about the greying of grey data?

Recent research by the European Broadcasting Union (EBU) on misinformation shows that only 59% of people in the European Union (EU) believe what they hear on the radio, 51% believe the television news, and only 47% believe what they read (Financial Times, 2018). Widespread fake news, misinformation, disinformation, spam emails, computer bots, botnets, web spiders, crawlers, and viruses erode our trust in the information and data we encounter in our daily lives, making trustworthiness a concern.

To further illustrate the concern of trustworthiness, consider that 269 billion emails are sent and received each day, of which 60% is spam. 56% of all internet traffic is from automated sources — hacking tools, scrapers and spammers, bots, and other malicious programs. Therefore, conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security are of increasing importance.

Uncovering deception and estimating the veracity of information and data is difficult now and will be even more so in the future.

Another factor impacting trust is the amount of data surrounding us. 2.5 exabytes of data are produced every day, the equivalent of 250,000 Libraries of Congress and 90% of all the data in the world that has been generated over the last two years. 13 million text messages are sent every minute, 4.4 million videos are watched on YouTube every minute and 1.7 megabytes of new information are created every second for each human being on the planet. Although the amount of information and data[1] around us is enormous, 99.5% of all data created is not currently being analysed and used. Still, we are hungry for information, as demonstrated by over 6.6 billion Google queries daily, 15% of which have never before been searched.

Grey literature

Various definitions of grey literature exist. The 12th International Conference on Grey Literature (GL12), held in Prague in 2010, defined it as "manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i. e., where publishing is not the primary activity of the producing body" (Farace, D. and Schoepfel, J., 2010).

Adams (2016) adds another twist to the definition of grey literature by proposing to look at it from the perspective of traditional publishing, which includes a peer-review process. Accordingly, grey literature is regarded as "the diverse and heterogeneous body of material

[1] Data is ‘facts or figures from which conclusions can be drawn’. Information is ‘data that have been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges’. www.statcan.gc.ca

that is made public outside, and not subject to traditional academic peer-review processes" (Adams at al., 2016).

The current definition faces some challenges, such as multiple types of originators — humans and machines — volume, and the speed of grey literature creation. It also faces the possibility of becoming obsolete due to its inability to differentiate between grey literature and other types of literature. Therefore, the following new definition was proposed: "***grey literature is any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process"*** (Savić, 2017).

This definition considers all major elements of the grey literature concept. Namely, long term preservation, sustainability, usability, and value, while acknowledging the lack of a traditional peer-review process for regular 'white' literature.

As Figure 1 shows, there are many types or forms of grey literature, although this paper mainly deals with grey data sets. The GreyNet website lists over 150 document types including databases, data sets, data sheets, data papers, satellite data, and product data.

Bibliographies	Rejected manuscripts	Publications from NGOs and consulting firms
Discussion papers	Un-submitted manuscripts	Videos
Newsletters	Conference abstracts	Wiki articles
PowerPoint presentations	Book chapters	Emails
Program evaluation reports	Personal correspondence	Blogs and social media
Technical notes	Newsletters	Data sets
Publications from governmental agencies	Informal communications	Committee reports
Reports to funding agencies	Census data	Working papers
Unpublished reports	Pre-prints	Company reports
Dissertations	Standards	Catalogues
Policy documents	Patents	Speeches
	Webinars	Reports on websites

Figure 1: Types of grey literature

There are many new sources of data, such as the Internet of Things (IoT), Machine to Machine communication (M2M), self-driven cars, robots, sensors, security systems, surveillance cameras, and many other systems or apps using AI and machine learning. The estimated number of currently connected electronic devices creating specific data varies by billions. Data produced by these devices is highly contextual and software dependent, making it hard to collect and process, and even harder to make sense of and preserve for future use.

White data

White data comes from the wider concept of 'white literature' (Jeffery, 2006), which is regarded as peer-reviewed and published literature, usually in the form of articles and books. It is often referred to, especially when talking about data, as open data. In other words, "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control" (Wikipedia).

The International Open Data Charter 2015 defines open data as digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere. It promotes the following principles:

- Open by default
- Timely and comprehensive
- Accessible and usable
- Comparable and interoperable
- For improved governance and citizen engagement
- For inclusive development and innovation

The European Union regards open (government) data as the information collected, produced or paid for by public bodies and can be freely used, modified, and shared by anyone for any purpose.

The US Federal Open Data Policy of 2013 refers to open data as publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. It is consistent with the principles of public, accessible, described, reusable, complete, and timely.

Many other countries such as Russia, China, and Japan also have their well-developed and defined national legislation and regulations regarding open data, particularly government and publicly funded data. It is interesting to note that the Russian Open Government Data (OGD) Recommendations of 2014, include requirements for licensing, mandatory procedures for data publication, rules for data publishing, data formats (CSV, XML, JSON, RDF), metadata format, and other technical requirements. Japanese regulations encourage the use of public information for both commercial and non-commercial purposes.

Grey data

Grey data represents a type of grey literature that maintains its basic facets, such as that it is recorded, that it is referable, sustainable, valuable, publicly available, and without a traditional peer-review. Just as with grey literature, it is this last characteristic that causes some data to be regarded as grey data. It is data that is useful and valuable, but not vetted by peer-review or other existing governance mechanisms.

Grey data is also an umbrella term that describes the vast array of data that organizations collect and use. It is often critical to an organization’s ability to innovate, enhance, and execute its core mission and it is usually collected for mandatory or compliance purposes, such as HR, budget and finance, contracts, procurement, facility management, registered library users, database subscriptions, and information collection maintenance. Besides being important for operational

and internal management, grey data is usually collected and managed for legal and regulatory purposes.

Many organizations that offer products and services collect data on users, product sales, and penetration of services not only for the purpose of production but also, importantly, for marketing.

Dark data

In contrast to white data, dark data is barely visible. It represents the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes, such as analytics, business relationships and direct monetizing.

Like dark matter in physics, dark data often comprises an organizations universe of information assets. Thus, organizations often retain dark data for compliance purposes only, without much further use. However, storing and securing data typically incurs more expense and sometimes represents considerably greater risk than value. Dark data is also recognized as a potential for revenue. According to Garner research, by 2020, 10% of organizations will have a business unit to make their data commercially available.

Data mining can be used to get value out of dark data as a data analysis process of sorting through large data sets to identify patterns, establish relationships, and solve problems. Data mining tools could even be helpful in allowing enterprises to predict future trends based on their historical data. However, some data archaeology might need to be performed to reuse preserved historical data by recovering information stored in formats that have become obsolete.

The table below offers an overview of major differences between white, grey and dark data.

Facet	White	Grey	Dark
Recorded	x	x	x
Valuable	x	x	x
Referable	x	x	x
Sustainable	x	x	
Used	x	x	
Public	x	x	
Peer reviewed	x		

Figure 2: Differences between white, grey and dark data

Grey data diversity

There is a great variety of approaches to grey data. A brief overview of the four most interesting and important approaches will be presented here. It includes:

- Data in the context of Grey System Theory
- Anonymous data as defined by the EU
- Unstructured data
- Unmanaged (risky) data

The Grey System Theory was made popular by Julong Deng (1982). He successfully developed a methodology which focused on the study of problems involving small samples and poor information, which is very often a common situation in decision- making, irrelevant of the field (e.g. economy, finance, politics and others).

In their book on grey data analysis, Sifeng Liu, Jeffrey Forrest, and Yingjie Yang (2017), present the fundamental methods, models and techniques for the practical application of grey data analysis. They also specifically talk about various types of data, e.g., black, white, and grey. For them 'black' indicates unknown, 'white' indicates completely known, and 'grey' indicates partially known and partially unknown information. More specifically, grey data represents small samples and poor information which is often only partially known, incomplete, inaccurate, and inadequate.

	Situation		
Concept	Black	Grey	White
From information	Unknown	Incomplete	Completely known
From appearance	Dark	Blurred	Clear
From processes	New	Changing	Old
From properties	Chaotic	Multivariate	Order
From methods	Negation	Change for better	Confirmation
From attitude	Letting go	Tolerant	Rigorous
From the outcomes	No solution	Multi-solutions	Unique solution

Figure 3: Black, white, and grey data according to the Grey System Theory

The European Union recognizes anonymous data as a type of grey data and presents it as a legal term used in the EU General Data Protection Regulation (GDPR). The reuse of personal data (processed for purposes beyond its original collection) is a key concern for the EU data protection law. GDPR applies only to information concerning an identified or identifiable natural person. Therefore, anonymized data is no longer considered to be personal and is thus outside the scope of GDPR, but there are problems with the techniques used to make data anonymous. The use of direct and indirect identifiers (quasi-identifiers) such as age, gender,

education, employment status, economic activity, marital status, mother tongue, and ethnic background, is of considerable concern.

GDPR regards pseudonymous data also as personal data, e.g. data which uses assigned IDs but where the research team has a key that can be used to connect the data to research participants. A process of well-planned and well-performed 'de-identification', or removal/editing of identifying information in a dataset to prevent the identification of specific cases is legally required. It should be carried out in such a manner that 'de-anonymization' is not possible, i.e. re-identification of data that is classified as anonymous by combining the data with information from other sources.

The European Union also insists on the principle of minimization. In other words, only the minimum amount of personal data necessary to accomplish a task/research should be collected, making most of the data, in fact, grey data.

Unstructured data represents any data that does not have a recognizable structure. It is data that, due to its non-existing structure, is not fit for use in a classical relational database. For example, text documents, email messages, PowerPoint presentations, survey responses, transcripts, posts from blogs, social media, images, AV files, machine and log files, and sensor data.

Due to the remarkable development of information technology tools such as AI, machine learning, deep learning, natural language processing, data mining, and predictive analytics, unstructured data is being analysed, categorized, classified, and efficiently stored. It should be noted that the line between structured and semi-structured data is very thin. By simply adding metadata tags to the data content, unstructured data can become semi-structured, or even fully-structured data.

Unmanaged or risky data represents, according to some estimates, almost 30% of corporate storage space. Another 30% of storage is filled with active data, while 40% of the data is inert and needs to be kept for archival or regulatory purposes. Out of the 30% of unmanaged data, 15% represents dark storage allocated but unused, 10% is orphaned data that should have been discarded long ago, and 5% is personal data that should not have been placed on corporate servers at all. To decrease the amount of unmanaged data, organizations need to implement well-established data governance policies that include relevant standards, life-cycle management guidelines, and compliance instructions, and quality control measures need to be put in place. Unmanaged data also represents considerable risk for organizations due to data clutter, liability issues, and increased security breaches, as well as the cost of maintenance, backups, disaster recovery, servers, space, electricity, and staff involvement.

Synthetic data is the newest and probably the most interesting part of the grey data spectrum. It is a specific set of data that is artificially manufactured, rather than directly measured and collected from real-world situations. Synthetic data is usually anonymized (stripped of identifying aspects such as names, emails, social security numbers and addresses) and created based on user-specified parameters resembling the properties of data from the real-world. It is an important tool to augment machine learning algorithms when real data is too expensive to collect, inaccessible due to privacy concerns, or incomplete.

AI systems that can learn from real data can also create data sets resembling authentic data. With further developments in information technology, it is expected that the gap between synthetic data and real data will diminish. Waymo LLC, a subsidiary of Alphabet Inc., tested its autonomous vehicles by driving 8 million miles on real roads and another 5 billion on simulated roadways — real proof of the power of synthetic data in practical life.

Unsettling grey data

In exploring grey data, it is also necessary to mention some of the challenges that can be unsettling Two areas of concern are the data itself and its basic purpose.

Concerns about the data itself come about because data is unverifiable, available for use but without any guarantee of its truthfulness. The danger in this lack of basic verification can result in inaccurate, or simply fake data, as is often demonstrated in news feeds. The structure of grey data is also often unclear, as well as its format and the tools required for analysis, making its use difficult. Encryption, an often misleading sign of trustworthiness, combined with redundancy, makes the use of grey data unsettling.

The purpose of data creation and its existence is another concern in ensuring the trustworthiness and usability of grey data. A good rule of thumb is to verify the source of the data and to be wary of questionable sources that may have clandestine reasons for creating the data in the first place, such as misinformation, defaming, or other hidden intents.

A final note regarding the unsettling use of grey data concerns the currently popular and widespread use of **F.A.I.R**. principles (**F**indable **A**ccessible **I**nteroperable **R**eusable). These principles are all valid and should be promoted and used; however, in my opinion, the most important one, **T**rustworthiness, is missing. **T**rustworthiness needs to be established, rigorously checked and followed.

Conclusions

Conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security have become important concerns. The trustworthiness of news and information, of grey and other literature types, and of grey data has become a public concern. The increasing amount of grey data being created impacts the way we process, disseminate, manage, and use this type of information; consequently, demanding greater trustworthiness.

Processing needs to be well-thought out and present from the beginning of grey data creation. Ad-hoc or post-processing can no longer be regarded as efficient. Environmental and technical, economic and financial, social and organizational constraints need to be taken into consideration for long-term grey data sustainability and usefulness. Its usability requires adequate IT tools, the availability of qualified human resources, and the protection of intellectual property and personal privacy.

To secure the future use and maintain the value of grey literature, intensive training, wide-spread cooperation, and proper management are needed. Only a small percent of businesses extracts the full value from the data they hold. The use of new IT tools such as AI, could improve its value, improve business results, bring measurable efficiency gains, and increase the quality of products and services. However, this will only happen if grey data reaches the required level of users' trust.

REFERENCES

Adams et al., 2016. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. International Journal of Management Reviews. 2016. (http://onlinel ibrary.wiley.com/doi/10.1111/ijmr.12102/full).

Farace, D. J. and Schoepfel, J. (Eds.), 2010. Grey Literature in Library and Information Studies. De Gruyter Saur, Germany.

Financial Times, 2018. Matters of Fact. News in the Digital Age, November 2018. Published by Financial Times in collaboration with Google.

GL12, 2010. Twelfth International Conference on Grey Literature. National Technical Library, Prague, Czech Republic. December 6-7, 2010. www.textrelease.com/gl12conference.html

Jeffery, K. 2006. Open Access: An Introduction. ERCIM News 64, January 2006.

Liu, S., Yang, Y., Forrest, J. 2017. Grey Data Analysis: Methods, Models and Applications. Springer.

Savić, D., 2017. Rethinking the Role of Grey Literature in the Fourth Industrial Revolution. 10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology. Available from: http://nrgl.techlib.cz/index.php/Proceedings. ISSN 2336-5021. Also published by TGJ (The Grey Journal) Special Winter Issue, Volume 14, 2018.