20th International Conference on Grey Literature

Research Data Fuels and Sustains Grey Literature

3-4 December 2018, New Orleans

WHEN IS ‘GREY’ TOO ‘GREY’?

A case of grey data

Dr. Dobrica Savić

linkedin.com/in/dobricasavic

Why are we concerned about the greying of grey data?

?

Amount of data

- 2.5 exabytes of data produced every day, equivalent to 250,000 Libraries of Congress
- 90% of all the data in the world has been generated over the last two years
- 13 million text messages sent every minute
- 4.4 million videos watched on YouTube every minute
- 1.7 megabytes of new information created every second for every human being on the planet
- 99.5% of all data created is not currently being analysed and used
- Over 6.6 billion Google queries daily, 15% never searched before

Trustworthiness

- Conformity to facts, accuracy, habitual truthfulness, authenticity, information source relability, security
- 269 billion emails sent and received each day - 60% is spam
- 56% of all internet traffic is from automated sources - hacking tools, scrapers and spammers, bots

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

- ✓ spam email
- ✓ fake news
- ✓ computer bots
- ✓ botnets
- ✓ web spiders
- ✓ crawlers
- ✓ viruses
- ✓ misinformation
- ✓ disinformation

Presentation at a Glance

- Grey literature definition
- Data types (White – Grey – Dark)
- Grey data
- Shades of grey data
- Synthetic data
- Unsettling grey
- Conclusions

- *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

Definition: Grey literature

> *GL represents any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process.*

The first recorded use of grey as a color in the English language was in AD 700.

Data types: White (open) data - Grey data - Dark data

Grey - between the extremes of black and white

Types of grey literature

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

- **New sources** of data and information: 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
- **Estimated number of currently connected devices** creating specific data varies by billions
- **Highly contextual and software dependent data and information** is hard to collect and process, and even harder to make sense of and preserve for future use

Fun Facts

The GreyNet website lists over 150 document types including databases, data sets, data sheets, data papers, satellite data, product data.

White (open) data

Wikipedia
The International Open Data Charter
The European Union
The US Federal Gov
Russia
China
Japan

The Open Government Data (OGD) Strategy sets forth the following basic principles:
-Government shall actively release public data
-Public data shall be released in machine-readable formats
-The use of public information shall be encouraged for both commercial and non-commercial purposes
-Specific measures shall be taken such as the prompt disclosure of public data that can be released, and results shall be steadily accumulated (**Japan 2012**)

The legal entities should establish a quality control system for scientific data to ensure the accuracy and usability of the data (**China 2018**)

rules for data publishing, data formats (CSV, XML, JSON, RDF), metadata format, and some other technical requirements. (**Russia**)

Open means anyone can freely access, use, modify, and share for any purpose, subject to the requirements that preserve provenance and openness.

Dark data

The information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (e.g. analytics, business relationships and direct monetizing).

Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only.

Storing and securing data typically incurs more expense (and sometimes greater risk) than value.

By 2020 10% of organizations will have a business unit for making their data commercially available.
(**Gartner**)

Data mining: the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends.

Data archaeology: preserves historical data using methods for recovering information stored in formats that are becoming (or have become) obsolete.

Grey data (GD)

- A type of GL that maintains its basic facets such as being recorded, referable, sustainable, valuable, publicly available, and without traditional peer-review
- An umbrella term that describes the vast array of data that organizations collect and use
- Useful and valuable data not vetted by peer review, or other governance mechanisms
- Often critical to an organization's ability to innovate, enhance, and execute its core mission
- Collected for mandatory or compliance purposes, such as HR, budget and finance, contracts, procurement, facility management, library users and collections
- Important for operational, internal management, and legal purposes
- Data on users, products and services collected for production or marketing purposes

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		

Grey or gray is an intermediate color between black and white. It is a neutral or achromatic color, meaning literally that it is a color 'without color' (Wikipedia)

Shades of grey

Unmanaged (risky) data

-30% of corporate storage space is filled with active data

-40% of the data is inert and needs to be kept for archival or regulatory purposes

-30% of the storage is used by unmanaged data (15% dark storage-allocated but unused; 10% orphaned data that should have been discarded long ago; 5% personal data which should not have been on corporate servers)

Risks (data clutter, liability, security breaches)

Cost (maintenance, backups, disaster recovery, servers, space, electricity)

Data governance (standards, life cycle management, compliance, quality control)

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

Synthetic data

- Artificially manufactured data rather than measured and collected from real-world situations.
- Usually anonymized (striped of the identifying aspects such as names, emails, social security numbers and addresses), and created based on the user-specified parameters resembling the properties of data from real-world.
- AI systems that can learn from real data can also create data sets resembling the authentic data. The gap between synthetic data and real data will diminish.
- An important tool to augment machine learning algorithms when real data is too expensive to collect, inaccessible due to privacy concerns, or incomplete.

Waymo (a subsidiary of Alphabet Inc.) tested its autonomous vehicles by driving **8 million** miles on real roads plus another **5 billion** on simulated roadways.

Unsettling grey

Warning signs!

Data

Unverifiable
Inaccurate (fake) data
Unclear structure
Difficult analysis (format, tools)
Encryption
Redundancy

Purpose

Questionable source
Misinformation
Hidden intent
Data abuse
Defaming

Findable
Accessible
Interoperable
Reusable

Fun Facts

The phrase "shades of grey" usually refers to a situation that is not clear, particularly with regard to whether or not something is categorically evil. When doubt comes into play, things are neither black, nor white, but are in a grey area (Martha Sorren, 2015)

Conclusions

- Increased amount of GD created will impact the way we process, disseminate, manage, and use it
- Increased number of GD types will demand higher trustworthiness
- Processing needs to be well-thought and present from the beginning of GL data creation. No ad-hoc or post-processing can be efficient
- Environmental and technical; economic and financial; social or organizational constraints need to be taken into consideration for long-term GD sustainability
- Usability of GL requires adequate IT tools, availability of qualified HR, protection of intellectual property, protection of personal privacy
- To secure future use and maintain the value of GL, intensive training, wide-spread cooperation, and proper management are needed
- Only a small percent of businesses extract full value from the data they hold. Use of new IT tools such as AI, might help get more value out of it, improve business results, bring measurable efficiency gains, increase quality of products and services

The real purpose of data is to uncover patterns, recognize correlations, and identify opportunities that translate to more efficient operations, smarter business decisions, and greater client satisfaction - ultimately, leading to higher profit margins.

It’s all grey until you find it!

Dean Giustini

Thank you!