GLOBAL MINING GUIDELINES GROUP

GMG

# GUIDELINE FOR SHARING OPEN DATA SETS IN MINING

PUBLISHED 2022

ARTIFICIAL INTELLIGENCE

Global Mining Guidelines Group (GMG)

## ABOUT GMG

The Global Mining Guidelines Group (GMG) is a network of representatives from mining companies, original equipment manufacturers (OEMs), original technology manufacturers (OTMs), research organizations and academics, consultants, regulators, and industry associations around the world who collaborate to tackle challenges facing our industry. GMG aims to accelerate the improvement of mining performance, safety, and sustainability by enabling the mining industry to collaborate and share expertise and lessons learned that result in the creation of guidelines, such as this one, that address common industry challenges.

Interested in participating or have feedback to share? GMG is an open platform, and everyone with interest and expertise in the subject matter covered can participate. Participants from GMG member companies have the opportunity to assume leadership roles. Please contact GMG at info@gmggroup.org for more information about participating or to provide feedback on this guideline.

GMG was formed out of the Surface Mining Association for Research and Technology (SMART) group as part of the Canadian Institute of Mining, Metallurgy and Petroleum (CIM) and with the support of other Global Mineral Professionals Alliance (GMPA) members.

GMG is an independent, industry-led organization.

## ABOUT GMG GUIDELINES

GMG guidelines are peer-reviewed documents that describe good practices, advise on the implementation and adoption of new technologies, and/or develop industry alignment. They are the product of industry-wide collaboration based on experience and lessons learned. The guidance aims to help readers identify key considerations, good practices, and questions to ask on the topic covered and enable operational improvements for safe, sustainable, and productive mines.

Once the guideline is reviewed and accepted by the project group steering committee, working group members peer review and GMG members within the working group vote to approve draft documents prior to their approval by the GMG Executive Council.

GMG guidelines are intended to provide general guidance only, recognizing that every situation will be different. Use of these guidelines is entirely voluntary and how they are applied is the responsibility of the user. These guidelines do not replace or alter standards or any other national, state, or local governmental statutes, laws, regulations, ordinances, or appropriate technical expertise and other requirements. While the guidelines are developed and reviewed by participants across the mining industry, they do not necessarily represent the views of all of the participating organizations and their accuracy and completeness are not guaranteed. See the disclaimer on p. iv for further detail.

## RELATED GMG DOCUMENTS

While guidelines are the primary output of GMG Working Groups, GMG also produces documents that supplement guidelines. These include:

- **White papers:** Educational documents that provide broad knowledge and identify further reading on a topic that is new to or not well-understood in the industry. These documents are reviewed throughout development and editing but do not undergo the working group review and voting process as guidelines do. These projects can lead to guideline development.
- **Reports:** Outcomes of outreach, industry research, and events can be presented in reports and can inform the priorities for developing industry guidance.
- **Landscapes:** Reviews of ongoing related work by other organizations on a key topic. These aim to provide the industry with an idea of what exists and prevent duplication of effort.
- **Case studies/other examples and tools:** These documents aim to share knowledge and provide examples for the benefit of the broader industry and supplement GMG guidelines.

## RELATIONSHIP TO STANDARDS

GMG guidelines are not standards and should not be treated as such. The guidelines can be used to assist the mining community with practices to improve their operations and/or implement new technologies. They aim to supplement, not replace, existing standards, regulations, and company policies. Guidelines can also be a first step in identifying common and successful practices and feed into standardization efforts. GMG does not develop standards but does participate in standardization efforts through partnerships.

## CREDITS

The following organizations and individuals were involved in the preparation of these guidelines at various stages including content definition, content generation, and review. Please note that the guidelines do not necessarily represent the views of the organizations listed below.

### Project Group

Open Data Sets for Artificial Intelligence in Mining

### Working Group

Artificial Intelligence

### Project Leaders

Louis-Pierre Campeau, Newtrax
Rob Johnston, CITIC Pacific Mining
Michelle Levesque, CanmetMINING

## PARTICIPATING ORGANIZATIONS INVOLVED IN THE PREPARATION OF THESE GUIDELINES

2BL, ABB, Accenture, Advisian, Afflatus, Agnico Eagle, Airth Solutions, Alex Atkins & Associates, Amazon Web Services, AngloGold Ashanti, Arundo Analytics, Atlas Iron, Ausdrill, Baatar Consulting, Balkan Gold, Barminco, BASF, BBA, Bixbyte, CITIC Pacific Mining, Clean Air Engineering, CMOC International, CNID, Colorado School of Mines, Connectmine, COREM, Curtin University, Dassault Systèmes, Decipher, Deloitte, Dendra Systems, Deswik, DetNet South Africa, DINGO, Edith Cowan University (ECU), EIPTET, Enaex, Endress + Hauser Group, Envirosuite, Epiroc, ERAMET – SLN, Exxaro Resources, First Majestic Silver Corp, Flow Partners, Geological Survey of South Australia, Glencore, Golder Associates, Groupe MISA, Hatch, Helios Consulting, Hexagon Mining, IBM, ICG, Ideon Technologies, Indstry4, IntelliSense.io, IO Solutions, Ivado, Kal Tire, KAZ Minerals Bozshakol, KJK, Koan Analytics, Komatsu, Life Cycle Geo, Lulea University of Technology, Maptek, Maxgeo, Mayhew Performance, Meglab, Metcom Technologies, METS Ignited, MGEOMET Services, Micromine, Minera Yanacocha, MineRP, Minerva Intelligence, MineSense Tech, MineWare, Mobilaris, Mpsa, MST Global, National Research Council Canada – IRAP, Natural Resources Canada, Nevada Gold Mines, Newmont, Newtrax, NIOSH, Nokia, Norilsk Nickel, Northern Star Resources Limited, Oceanagold, Off World, Optika Solutions, Oracle, Orica, OZ Minerals, Pacific GeoTech Systems, PETRA Data Science, Pivot Industries Limited, Precision Mining Solutions, Queen's University, RAK Developments, Read Eagle, ReRisk, Resolution Systems, Rio Tinto, Rithmik Solutions, Rockwell Automation, Roy Hill, Rudplan DOOEL, RWTH Aachen University, Sandvik, Schneider Electric, Seequent, Semafo, SMART Systems Group, Solvay, Spectris Advance Mining, SPIE Plexal, Spring Tech, SRA Information, SRK Consulting, Swann, Symbiotic Innovations, Tanuki Service, Teck, Tellus Mining, The Doe Run Company, UAE Tetouan Maroc, UFRJ, Underground Mining Solutions, Universite Laval, University of Alberta, University of Queensland, University of Western Australia, VDMA, VMWare, Wabtec Digital Mine, Wenco, Wipro Consulting, Worley, and X-Analytics.

## PUBLICATION INFORMATION

Guideline Number: GMG13-AI-2022
Published: 2022-04-21
Revision Cycle: 2 years

## DOCUMENT USAGE NOTICE



## CONTACT INFORMATION

**Global Mining Guidelines Group**
info@gmggroup.org
gmggroup.org

## DISCLAIMER

This publication contains general guidance only and does not replace or alter requirements of any national, state, or local governmental statutes, laws, regulations, ordinances, or appropriate technical expertise and other requirements. Although reasonable precautions have been taken to verify the information contained in this publication as of the date of publication, it is being distributed without warranty of any kind, either express or implied. This document has been prepared with the input of various Global Mining Guidelines Group (GMG) members and other participants from the industry, but the guidelines do not necessarily represent the views of GMG and the organizations involved in the preparation of these guidelines. Use of GMG guidelines is entirely voluntary. The responsibility for the interpretation and use of this publication lies with the user (who should not assume that it is error-free or that it will be suitable for the user's purpose). GMG and the organizations involved in the preparation of these guidelines assume no responsibility whatsoever for errors or omissions in this publication or in other source materials that are referenced by this publication, and expressly disclaim the same. GMG expressly disclaims any responsibility related to determination or implementation of any management practice. In no event shall GMG (including its members, partners, staff, contributors, reviewers, or editors to this publication) be liable for damages or losses of any kind, however arising, from the use of or reliance on this document, or implementation of any plan, policy, guidance, or decision, or the like, based on this general guidance. GMG (including its members, partners, staff, contributors, reviewers, or editors to this publication) also disclaims any liability of any nature whatsoever, whether under equity, common law, tort, contract, estoppel, negligence, strict liability, or any other theory, for any direct, incidental, special, punitive, consequential, or indirect damages arising from or related to the use of or reliance on this document. GMG (including its members, partners, staff, contributors, reviewers, or editors to this publication) is not responsible for, and make no representation(s) about, the content or reliability of linked websites, and linking should not be taken as endorsement of any kind. We have no control over the availability of linked pages and accept no responsibility for them. The mention of specific entities, individuals, source materials, trade names, or commercial processes in this publication does not constitute endorsement by GMG (including its members, partners, staff, contributors, reviewers, or editors to this publication). In addition, the designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of GMG (including its members, partners, staff, contributors, reviewers, or editors to this publication) on the legal status of any country, territory, city, or area or of its authorities, or concerning delimitation of any frontiers or boundaries. This disclaimer should be construed in accordance with the laws of Canada.

## EXECUTIVE SUMMARY

As technology advances, data can provide opportunities to solve problems in various areas including accelerated research, increased transparency, and identification of novel solutions to problems. Unfortunately, the appropriate data are not always readily available.

Open data is defined as digital information that is made available with as few technical or legal restrictions as possible so that it can be freely shared, used, interpreted, and built upon anywhere by anyone. This definition is paraphrased from the Open Data Handbook (Open Knowledge Foundation, 2020) and Government of Canada's "Open Data 101", (2020). In the context of this guideline, open data refers to machine-readable digital data.

The purpose of this guideline is to provide mining industry stakeholders with best practices for data sharing that are based on existing initiatives so that they can benefit from the opportunities that open data can offer. This guideline is directed towards readers who intend to share data with others, those who are involved in the approvals process, and users who want to use open data shared by the mining industry.

### Management Considerations

A data license is typically used before sharing and publishing data to outline the data providers' intended use while giving them protection. It also provides clarity to the data consumer, preventing them from potentially infringing the rights of the owners. Different types of licenses are available for different purposes. License types can typically be divided into open (without technical or legal restrictions), non-commercial, partially open or restricted usage, and closed. Existing frameworks such as Creative Commons and the Montreal Data License can be used to cover general requirements.

Sharing data provides benefits, which include supporting innovation and research and allowing the public access to information to help improve decision-making in operations. Before implementation, addressing the challenges of cost, legal issues, storage, privacy, and common language associated with collection, administration, internal communication, and maintenance of open data is crucial to minimize the challenges and maximize the benefits of sharing the data.

### Implementation Considerations

Identifying what data should and should not be shared is very important before implementation. The data set should be well-documented, reliable, usable, accurate, relevant, and in an accessible format. If a data set is commercially sensitive, contains personally identifiable information (PII) or sensitive data, or poses a security risk, sharing the data sets should be avoided unless these risks can be mitigated. A risk assessment should be completed based on the organization's policies and risk tolerances.

When making a data set open, it should be submitted in a machine-readable format that is open and logical. If possible, any community consensus on format or formats of existing data should be prioritized. It is also important to identify the appropriate anonymization requirements and techniques.

It is recommended that a formal approval process is adopted when releasing data. The documentation provided for approval to release data typically includes information that provides an overview of the original data and its structure, a description of anonymization procedures, an overview of the resulting data, and attestation or "sign-off" from key stakeholders that the data set is acceptable to share. Selecting the appropriate hosting and listing platform is the final step before making the data set open.

## TABLE OF CONTENTS

# 1. INTRODUCTION

In the transition to a data-driven world, data can provide opportunities to solve problems in various areas, for example, accelerated research, increased transparency, and identification of novel solutions to problems. However, these data are not always available.

Open data is defined as digital information that is made available with as few technical or legal restrictions as possible so that it can be freely shared, used, interpreted, and built upon anywhere by anyone. This definition is paraphrased from the Open Data Handbook, (Open Knowledge Foundation, 2020) and the Government of Canada's "Open Data 101", (2020). In the context of this guideline, open data refers to machine-readable digital data.

The purpose of this guideline is to provide mining industry stakeholders with best practices for data sharing that are based on existing initiatives so that they can benefit from the opportunities that open data can offer. The document is designed to support:

- Anyone in a mining organization who intends to share data with others
- Anyone in the mining industry who approves sharing data between entities
- Anyone who wants to use open data shared by the mining industry

This guideline is divided into the following two sections:

- Management considerations that explain the value of open data sets and provide the context required to understand the process and help the guideline user decide whether sharing open data sets is appropriate for their situation (Section 2)
- Implementation considerations that explain processes for sharing open data sets once the guideline user has decided to proceed (Section 3)

**Related Publication**

- Foundations of AI: A Framework for AI in Mining (GMG, 2019): This white paper is a primer on AI in mining and presents general information that relates to data.

Guidance on open data curation and management are outside of the scope of this guideline. Detailed guidance on ownership of open data and interoperability, while noted, are also out of the scope of this guideline.

# 2. MANAGEMENT CONSIDERATIONS

This section describes the potential custodians, provides guidance on license types, describes some of the benefits of sharing data, and explores the overall risks and costs associated with open data. The considerations offered in this section are intended to provide the context required to understand the process and help the guideline user decide if open data sets are appropriate for their situation.

## 2.1 Custodians of Open Data

Custodians of mining-related open data can be:

- Mining companies (including exploration and production)
- Mining equipment, technology, and services organizations
- Government agencies
- Non-governmental organizations
- Academic and research institutions

These data sets can vary by date, structure, accuracy, and data sources. However, if combining data sets from various sources, there can be an impact on interoperability.

## 2.2 License Types

When publishing a data set, it is important to supply a data license with it in order to outline the terms of use for the data. The license has two main purposes:

- It outlines the data provider's intended use of the data and provides them with some protection.

- It gives clarity to the data consumer regarding how they can use the data without infringing on the rights of the owner.

In general, data licenses can apply to a few distinct categories:

- **Open:** A data set may be released with very few restrictions on its intended use. There can be some minimal requirements such as attribution, but any consumer is free to use the data by whatever means they see fit.
- **Non-Commercial:** Data sets may be published openly but have significant restrictions in place. For example, the data set may be available for academic use but cannot be used commercially.
- **Partially open or restricted usage:** Data may be made available to a small group of entities, and each entity must agree to a custom agreement. For example, a group of data producers may agree to open data between each other on a like-for-like basis (i.e., each must contribute an equal amount of data to the common group).
- **Closed:** Closed data sets are those that may be shared between two parties for a very narrow purpose. They are typically governed by non-disclosure agreements (NDAs).

Considerations when choosing a data license include the ownership of the data. The data owner could be the mining company, OEM, or a third-party company developing or creating a commercialization opportunity. For example, the licensing requirements may differ especially when it comes to commercialization if it is owned by an original equipment manufacturer (OEM) or a third party.

Some frameworks exist that can help with licensing. A general licensing framework, such as Creative Commons, aims to provide licenses that cover general requirements. For example, the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license allows the user to adapt the works in any non-commercial way as long as they give attribution (Creative Commons, 2020).

While standard frameworks like those from the Creative Commons are meant to cover general use cases, more specific frameworks can provide further guidance for AI and machine learning use cases. One such example is the Montreal Data License (described in Benjamin et al., 2019), which provides guidance about the release of data, clarity for individuals and companies that make data available to third parties, and a framework for data licensing.

Benjamin et al. (2019) envision the use of data in AI and machine learning as an incremental process that spans the value chain from extraction to output; this data may be used in different ways and for different purposes. They have developed standard definitions around the data itself and standardized definitions for accessing, labelling, distributing, and representing data. They also define the use of data in conjunction with models, summarized as follows:

- **Benchmark:** With or without a trained model
  - With training a model: Models can be trained on the data set for the benchmarking purpose only
  - Without training a model: No models can be trained on the data set
- **Academic usage**
  - Research: Can use the data in a non-commercial environment for analysis
  - Publish: Defined as per research, but also includes the ability to publish models to a third-party
- **Commercial usage**
  - Internal use: Can use the data in a commercial environment for internal use only
  - Output commercialization: The same as internal use, but the output of the model can be used in products
  - Model commercialization: The same as output commercialization, but the model itself can be published to third parties

Table 1 provides an example table from the Montreal Data License framework that summarizes the rights granted with models.

There are different types of licenses that are available for different purposes. Choosing the correct one when sharing data is important for protecting the data as well as avoiding inadvertent use of the data. Benjamin et al. (2019) provides one example of an organization that has published and developed license types and is described here because it applies specifically to machine learning data sets, but others exist, and it is up to the user to select the most appropriate framework for their situation.

**Table 1. Montreal Data License Summary of Rights Granted in Conjunction with Models**

| | Reuse of Trained Model | Use of Output | Trained Model Made Accessible to Third Parties | Notes |
|---|---|---|---|---|
| Benchmark | × | × | × | |
| Research | ✓ | ✓[1] | × | [1] Use of output or the trained model must be tied to research purposes (i.e., reuse rights do not extend beyond research rights granted). |
| Publish | ✓ | ✓[2] | ✓[2] | [2] Use of output or the trained model must be tied to research and publication purposes (i.e., reuse rights do not extend beyond research rights granted). |
| Internal Use | ✓ | ✓[3] | × | [3] Output can be used internally for any purpose but cannot be made available to third parties. Internal use excludes output commercialization and model commercialization. |
| Output Commercialization | ✓ | ✓ | × | Output commercialization would allow SaaS business models in which output is made available to third parties. |
| Model Commercialization | ✓ | ✓ | ✓ | The trained model itself may be made available to a third party, with or without the output. |

Source: Reprinted from Benjamin et al., 2019, p. 13.

Regardless, when creating a license, it is recommended to base it on an existing framework, which may require professional consultation prior to sharing data.

Benjamin et al. (2019) also provide several tools to assist those sharing data. Table 2 is a top sheet (Benjamin et al., 2019) for describing the data. They also provide licensing language that can be used under CC-BY4 for AI and machine learning use cases (Benjamin et al., Appendix 4).

**Table 2. Top Sheet for Licensed Rights**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Licensor** | Name/Corporate Information of Licensor | | | | | |
| **License Data Set** | Description of Licensed Data Set | | | | | |
| **Technical Specifications** | Data Set Size, Format, and Other Technical Specifications | | | | | |
| **Rights to Data (Stand-Alone)** | Access | Tagging | | Distribute | Re-represent | |
| **Rights to Data in Conjunction with Models** | Benchmark | Research | Publish | Internal Use | Output Commercialization | Model Commercialization |
| **Credit/Attribution Notice** | | | | | | |
| **Designated Third Parties** | | | | | | |
| **Additional Conditions** | | | | | | |

Source: Reprinted from Benjamin et al., 2019, p. 14.

## 2.3 Benefits of Sharing Data

Sharing data can provide many benefits, from supporting innovation and research to enabling industry benchmarking. The Government of Canada's Open Data 101 (2020) also provides a quick reference that summarizes the general benefits of open data.

### 2.3.1 Research, Innovation, and Testing

The use of open data sets can help stimulate the innovation environment and help new, innovative solutions become available faster. Mining companies have the data while the technology companies have the innovations. Unless the innovation can be trained or tested on the data, it cannot be proven, and open data sets can provide a solution to this problem by connecting them. Scaling can also be a challenge faced by technology subject matter experts if the required resources to collect the data needed to develop and refine their technologies are not available. Open data sets can provide opportunities for them to test their technology at scale.

Additionally, open data sets can share important challenges between the mining community and academic institutions so that those challenges and the characterized data sets can be explored by several different specialists, using different algorithms and methodologies, resulting in improved solutions.

### 2.3.2 Industry Benchmarking and Comparison

Open data sets enable benchmarking and comparisons and help mining companies and technology providers make decisions and glean insights that are based on industry-wide data rather than data limited to a single operation or company. Using open data sets, technology companies can check the validity and applicability of their solutions against a known example. These direct comparisons can help the mining companies determine which solution is best for their needs.

More broadly, benchmarking using open data sets can help to provide an enhanced understanding of industry trends and patterns or the ability to better predict outcomes. One example in the mining industry is the tailings dam data stored at the Global Tailings Portal. These data are shared with the goal of reducing risks caused by tailings dam failures.

### 2.3.3 Broader Collective Benefits

Overall, a culture around open data sets can also offer collective benefits for the industry, for example:

- Helping the mining industry establish a culture that values transparency and optimization that can strengthen relationships based on shared industry knowledge
- Attracting a data-oriented workforce to the mining industry and training them in specific mining challenges, as many experts in the field support open collaboration
- Establishing a common interest with the top technology industries in the world
- Enabling common approaches to safety
- Encouraging industry standardization and enhanced interoperability

### 2.3.4 General Benefits of Data Sets

While the more general benefits of having data sets do not require data sets to be open, they can be enhanced by the benefits open data sets offer when it comes to innovation and benchmarking as described above. The following are examples of ways in which data sets can be used to improve the overall mining operational environment:

- Open data sets that include hazardous and potentially hazardous situations can be used to train AI or machine learning algorithms to identify trends and lead to better alerting/prediction of those situations in the future.
- Asset health data gathered from mobile equipment such as tire pressure and engine performance could improve failure predictions and lower costs of unplanned maintenance.
- Historical data on the environment could be used to help underground mines make decisions to improve worker safety and reduce waste.
- A large and diverse open data set that comprises sensors (e.g., optical, LIDAR, radar, IMU, GPS, etc.) of different scenarios in a mining environment can assist the developers of autonomous mobile equipment tailor their solutions for mining.
- Data sets can contribute to more consistent data across operations, which can help improve supply chain management.
- Data sets can help to enable valuable internal communications between those with data science or domain expertise and operational or business professionals, allowing people with different skill sets and perspectives to take a fresh view on the data to help uncover solutions.

- Open data sets can spur new innovation by allowing a wider group of participants to identify potential issues and/or opportunities and relate them to possible technological solutions that operators are currently not aware of.

## 2.4 Challenges, Risks, and Costs

There are costs, challenges, and risks associated with the collection, administration, internal communication, and maintenance of any open data set. When choosing whether or not to make an open data set available for use, it is important to weigh potential future value generation against the cost or risks.

The risk tolerance can also depend on the setting in which the data are being used. See also Section 3.3 for more specific guidance on the risk assessment process.

### 2.4.1 Effort and Cost

The process of collecting and curating data always has an associated cost, particularly with larger data sets. Storage and distribution costs can be significant if part of the license requires maintaining a degree of controlling and tracking over who has access to the data. However, if data types, formats, data gathering processes, and storage repositories are well-defined, the costs associated with sharing data can be reduced. Curating data may not necessarily be required, other than sanitizing column headers using a consistent convention or data dictionary.

Raw data might be more beneficial than pre-processed/curated data. If a data producer shares data that is as raw as possible, they can avoid the effort and cost spent grooming data before it is released. Such a process would allow:

- Reduced cost for sharing
- Data scientists' ability to explore perceived outliers, which can be a valuable source of information

However, if the raw data are sensitive to the mining operation in any way, then it can be transformed to make it less sensitive (e.g., averaged, scaled).

Some considerations for making a data set that can be open include:

- Deciding whether to proceed with a dedicated effort to collect data to avoid capturing information that there is no intention to share, or to proceed with collecting data as part of a normal operation where the objective is to triage or filter the data after it has been collected and before it is shared
- Determining how to share data, for example if it can be shared online or if it requires a hard drive or other mechanism to accommodate the size
- Determining if the data should be tracked once it has been shared and if so, making sure it has been appropriately resourced

### 2.4.2 Legal Issues

Stakeholders may avoid sharing data because it is easier to simply avoid all the potential legal implications rather than think through ways to share data without running into legal issues. A checklist/flow chart like the data risk assessment tool (DRAT) process can be a useful tool (Sikorska & Hodkiewicz, 2019).

### 2.4.3 Storage

While storage depends on the size of the data set, these data sets have no common repository or aggregator for storage. See the list of existing open data set platforms and examples on the GMG website, which lists some of the storage options that are available.

### 2.4.4 Privacy

Open data sets in mining should not contain personally identifiable information. The DRAT process provides some examples of how to deal with personally identifiable information (Sikorska & Hodkiewicz, 2019).

Commercially sensitive information, a subset of privacy, is related to the commercial aspects of the business. The management of commercially sensitive information can be ameliorated, for example, by releasing data with a time lag (i.e., after public reporting intervals).

#### 2.4.5 Variation

It is not uncommon for there to be variation and segregation in data collection and management between different operations, even within one company. It is important to be aware of the challenges associated with these issues and consider ways of mitigating them because they can affect the quality of the data sets.

## 3. IMPLEMENTATION CONSIDERATIONS

This section covers the processes for sharing open data sets. It offers guidance on what data should and should not be shared as well as processes for preparing and extracting data, conducting a risk assessment, obtaining approvals, and making data sets open.

### 3.1 Identifying Data to be Shared

Not all data can or should be publicly shared. There are many aspects to keep in mind when evaluating whether or not to make data publicly available.

Ideally, the published data sets should be:

- Well-documented or self-documented
- Reliable and accurate, with metadata included
- Usable
- Accessible in a well-known format

The publisher should make sure that published data does not include:

- Data that could lead to insider trading (special consideration should be given to data regarding productions, demand, and sales)
- Data that contain personally identifiable information (e.g., social security number, address, phone number, financial information)
- Data that contain information that might pose a security or confidentiality concern
- Commercially or operationally sensitive data

Those sharing data should take extra care to make sure that, even with data anonymization (Section 3.2.2), cross-referencing the data with other data sources does not reveal or re-identify personal information (for further detail on re-identification, see Lubarsky, 2017). Also, regulations regarding privacy can vary, depending on the region in which the data are published.

For better discovery, those sharing data should consider categorizing them according to well-established mining subjects. Some examples include:

- Exploration
- Topological and geological
- Environmental
- Socio-economic
- Market information (prices, exchange, interest and inflation rates, opportunity cost, etc.)
- Mining and extraction
- Equipment monitoring and health
- Process and transformation
- Health and safety

### 3.2 Process for Extracting and Preparing Data

This section outlines some considerations about standard data formats and anonymization, which are key parts of the process for extracting and preparing data. While this section does not go into detail, specific published examples can be a good resource for providing that detail. Some examples include:

- Global-scale remote sensing of mine areas (Werner et al., 2019)

- Analysis of Strainbursts in the Sudbury Region and Numerical Modelling of Destress Blasting (Gingras Little et al., 2017)

Some further examples of platforms and data sets compiled by GMG are available here.

### 3.2.1 Standard Formats

No specific formats are required to submit a data set, as long as the data are machine-readable. The only restrictions include:

- The format should be open.
- Parsing instructions should be provided with the data and in a separate README file.
- Any compressed files should not use a proprietary compression format or encryption.

Proprietary formats, such as.dwg, should be avoided as much as possible since they restrict the data access to specific software users. If no other options are available, these formats are better than nothing. Most of the time, open alternatives to proprietary formats are available within software-reading proprietary formats (e.g., saving in.dxf rather than.dwg). In addition to a README file that explains how to parse the file, those submitting the data should consider if a parser for any binary data format is necessary.

In general, if a community reaches a consensus about a specific format for the type of data to share, that format should always be preferred. Also, if entries of the same data type already exist, the same format that prior data sets used should be prioritized as much as possible. This prioritization should ease the process of grouping different data sets together for larger analysis.

### 3.2.2 Anonymization

Anonymization can be required for many reasons and is not limited to the inclusion of personal or competitive information. The simplest way to anonymize is to avoid sharing some data, but this tactic depletes the value of the data set. Since anonymization often leads to a loss of information and value, it is important to ask if anonymization is really required. The process of anonymizing a data set removes any data that carries the most information about its source. Instead of trying a one-size-fits-all anonymization process, each data set should be considered individually to establish what needs to be anonymous.

This section does not cover specific anonymization techniques, as they are generally applicable and are not specific to mining, but many guides can be found online that describe the different methods (e.g., see the references for a guide to basic data anonymization techniques from the Personal Data Protection Commission Singapore, 2018).

Differential privacy can be a consideration as part of the anonymization process. Differential privacy refers to a way to share data set information with the public without revealing information about those people within the data set. It allows those sharing data to detail group patterns inside the data set. However, there are trade-offs to be considered. The Harvard University Privacy Tools Project (2021) provides further information and tools.

Table 3 provides some considerations that are especially pertinent to the mining industry. Please note that these considerations are based on industry experience and may not be applicable in all situations. They do not replace regulations, standards, or company policies.

Those making data anonymous should follow internal guidelines that have been established to protect the organization's data and should be agreed upon by internal stakeholders including those working in cybersecurity, human resources, legal, operations, and management.

**Table 3. Anonymization Considerations**

| Consideration | Description | Why it is important |
|---|---|---|
| Names and IDs | Anything that has a name should be replaced by a unique ID, whether it is a person's name, a machine name, or a location. This allows the data set to maintain the relationship between different entries without disclosing information that could be linked to a specific source. | Replacing a name with an ID is necessary because most mining companies can have their own standards for naming locations or equipment, so the original name could be traced back to a specific company. |
| Dates and times | The seasonality in the data set is an important consideration when adjusting timestamps, and only adjustments that would not contradict it should be made. For example, anything that is affected by weather and the four seasons should be shifted by values close to integer years so that the seasons are coherent in the values. | Dates and times can be changed to remove time-oriented information. For example, timestamps for productivity data could be adjusted by a few hours in order to hide the real shift start time, which could be used to identify the source of data. |
| Mine plans/maps | Mine maps and plans are nearly impossible to completely anonymize. While having a profile of the terrain can add a lot of value to the data set, it is important to be careful about keeping the information general. | For underground mines, a lot of information can be extracted from the plan, for example, the mining method, type of orebody, and a rough estimate of yearly production. Furthermore, even if some characteristics are common to many mines, each mine plan is unique and should be easily identifiable by anyone who previously worked there or accessed the plan. |
| Productivity data | Productivity data, such as the productivity data from the equipment and mill, can be very sensitive. Therefore, it is important to be very cautious when sharing such information. In general, only the total productivity needs to be hidden, which allows part of the information to be shared. In such cases, a logical splitting point should be used to decide what to share and what to hide. | Removing some data is a good way to avoid tracing it back to a specific location, for example, by sharing data for one specific model of truck and not for others. If the data set includes data from 50 trucks, and only one mine in the area has that many trucks, sharing only 20 of them should help in preventing identification. |
| Already available data | When anonymizing, it is important to assess not only the data that are about to be shared but also the data that are already available online and how they could be traced back to the anonymized data set. | Even if the anonymized data cannot be retraced to a specific mine or person, its combination with or comparison to another available data set might make it identifiable. |

## 3.3 Risk Assessment

This section offers guidance for completing the risk assessment before making data open. Please note, that this is only intended to offer good practices and it remains the responsibility of each organization to determine their course of action based on policies set and their respective risk tolerance levels.

When addressing the topic of legal or security aspects relating to the data or information sharing in open sets or formats, it is good practice to incorporate the following three considerations into your risk assessment or risk management process:

- Can the organization maintain the necessary levels of compliance, abiding by and aligning with internal and external rules, ethical codes, regulatory aspects, and confidentiality and contractual/commercial clauses and requirements?

- Can the organization enforce consistency in future decisions, processes, and procedures on a case-by-case basis? Can you do this with the same set of rules, standards, and guiding decisions to maintain consistent review and consideration of legal ramifications and impacts before you make a decision (not having multiple or contradicted outcomes for similar issues)?
- Are the chosen controls, processes, or procedures associated with risk and decision-making complete in nature? Do the identified controls consider obligation and enforcement as well as whether and how such terms can be satisfied or applied as they pertain to legal and security-related aspects?

The DRAT process and ISO 31000, summarized in the following subsections, are two examples of tools that can be used to complete the risk assessment.

### 3.3.1 DRAT Process

The DRAT process can be useful for completing the risk assessment; an in-depth paper and flow chart for reference is available at https://drat-process.com/home.

The purpose of the tool is to build a consistent way to release asset-related data. The tool assesses and offers suggestions regarding controls to help you manage risks associated with the release of engineering asset management (EAM) datasets. Some considerations include where organizations place approval accountability and manager considerations in approving a release. The tool helps you make sure, for example, that you have based recommended restrictions and controls on the actual risk a data set poses and that you suitably manage the data owner's needs for confidentiality.

The DRAT flow chart asks guiding questions, for example:

- Have these assets, processes, or procedures been associated with known bad events?
- Is there a greater good to the company or community from analyzing these specific assets, processes, or procedures?
- Would the publication of this sensitive data have more than medium-level consequences as per your risk assessment process?
- Can the data set be pruned to exclude the most recent data and still be useful for the research intent?
- Will the highly sensitive data be used with similar data from other companies?
- Does the data set in its rawest available form contain personal data about people?

### 3.3.2 General Considerations Mapped to the ISO 31000 Process

Table 4 maps some general considerations about open data sets to the structure outlined in ISO 31000 (ISO, 2018).

**Table 4. General Considerations for the Risk Assessment**

| Step | Considerations |
|---|---|
| Establish the context | The types of risks to consider in this context are typically legal and security risks associated with open data sets and sharing of data. It is also worth considering the risks associated with the costs incurred throughout the process (e.g., maintaining a replica database, approved connection and firewalls). |
| Identify the risk | Some potential questions to ask in order to identify the risks of sharing open data sets include:<br>– What data will be shared?<br>– What is the content and specific information associated with the data?<br>– Would sharing the data create any breach of trust, policies, or regulatory aspects?<br>– Would sharing the data create reputational challenges?<br>– Do different strategies or platforms exist to share the data safely (conditions or license agreements)?<br>– Would sharing the data create any unfair or illegal trade or investment advantage to the market?<br>– Would sharing the data compromise the organization's integrity (perceptually or operationally)?<br>– Would sharing the data make the organization vulnerable to attacks, infiltration, or hackers? |
| Analyze the risk | – Focus on identifying existing controls within the organization that can eliminate, manage, or reduce (mitigate) the risk.<br>– Consider the likelihood/probability and the potential consequence of the risk. For example, if the data or information to be shared contains confidential information that is illegal to share, the consequences may be significant due to legal action, reputational impact, share-price impact. |
| Evaluate the risk and decide | Review existing guides for treating particular risks. Many legal and security-related risks have requirements prescribed internally or in relevant legislation, standards, codes, and other external requirements. Options include:<br>– Avoiding the risk by deciding not to share the data. This should remove possibilities of harm but can also often eliminate the potential opportunities from sharing the data.<br>– Changing the likelihood and/or the consequences of the risk.<br>– Both of the above options. |
| Monitor and review | – Determine if any conditions change due to changes in external environments.<br>– Scan and assess the external landscape for changes, including regulatory and license aspects and legal and security-related legislation as it pertains to data and information<br>– Incorporate lessons learned and identify areas for improvement. |
| Communicate and consult | Throughout the process, consider the following questions:<br>– What stakeholders need to be consulted at which stage of the process?<br>– Are the expectations clear?<br>– How will you address different views? |

## 3.4 Generic Process for Approval to Release Data

Table 5 provides an example of information needed in a typical document for approval to release data.

**Table 5. Example Structure for Approval Document**

| Section | Information |
|---|---|
| Overview of original data | – Source of data<br>– Executive summary of the data contained<br>– Ownership of the data<br>– Original attributes description (describe each column's content)<br>– List of sensitive information contained in the data set and impact if leaked |
| Anonymization | – Description of anonymization procedures<br>– List of sensitive information as well as which anonymization method removes the risk and how |
| Overview of resulting data | – Anonymized attributes description (describe each column's content)<br>– Potential use<br>– Benefits of sharing<br>– Risk evaluation<br>– Sharing method/location<br>– Frequency of review of content |
| Sign-off | From technical experts, SMEs, and executive management to attest that the sensitive information has correctly been listed and assessed, that the anonymization procedure is adequate to prevent risk, and that it is acceptable to share |
| Follow-up review | Including dates, signatures, notes of modifications, and other comments |

## 3.5 Process for Making Data Sets Open

Once the decision has been made to release open data, the selected data sets need to be hosted online and information shared about how to access it. While in some cases the data set will be made publicly discoverable, if the data is to be released on request, information is typically made public that explains the data set and provides information about how to request access.

There are multiple options to make data sets open, including:

- Cloud vendors
- Open data providers

There are two key considerations if making the data set widely discoverable is desired:

1. **Hosting** is the ability to provision the required infrastructure to store and manage access to data. Most cloud vendors have the ability to host data for a fee and make it public. Sometimes, depending on the data size, cloud vendor assessment, and policy, these vendors provide free hosting and listing services. Organizations also have the option to host data on their infrastructure and list them on public sites to enable discovery.
2. **Listing** is the ability to provide a list of open data sets on the web with links to the hosted data. Generally, the hosting provider also provides a listing of the data set. However, once data are hosted, the listing of data sets should occur on multiple trusted sources to make sure that wider public discoverability and usage are possible.

For those making data available on request or might want to make them discoverable only to a targeted audience, considerations about hosting and listing will be similar but might need to be more tailored to the data set and its audience and accessibility model.

One key consideration when making data sets open is their metadata information, which describes the data. There are multiple standards available on metadata (see, for example, the Research Data Alliance Metadata Standards Directory). The document describing the data set should typically include the following information:

- What sensors were used and what the performance characteristics were (e.g., noise, error)
- If any calibration was done
- If measurements are raw or interpreted
- If there is a configuration (e.g., if the sensors are mounted on a vehicle, how they are mounted)
- If there are any known problems or characteristics in the data set that the user should consider

However, most of the time, guidelines for listing data sets are typically provided by the hosting vendor (see the list of open data sets and platforms compiled on the GMG website).

### 3.5.1 Curation and Hosting Considerations

It is also important to consider how to curate and host the data before deciding where to release the data. Some key questions to ask to make sure the process is effective include:

- How much will curation and hosting cost?
- What is the reputation of the hosting site?
- How sustainable is the site (i.e., how long will the data be available on the site)?
- Will the data set be updated, and if so, how often?
- What are the policies and services of the hosting site?
- Is data structured in a way that can facilitate analysis?
- Are any actions required to make sure that data remains valid and usable?
- How secure is the hosting site?

## 4. FUTURE WORK

This guideline covered the process of sharing open data sets, both from a management and implementation perspective. The next step on this guideline is to gather input and experiences of guideline users that can be used to share case studies and examples of the open data sharing process and to improve future editions of the guideline. Further work might also involve updating the resource that identifies existing data sets and platforms, a preliminary list available here.

Potential future work on the topic would cover other considerations around open data sets that were either not covered or not covered in detail in this guideline. These include:

- Open data set curation and management
- Data ownership as it applies to open data sets

The decision-making and prioritization of future work on the topic will be completed by the GMG Artificial Intelligence Working Group. If not specific to AI topics, some further work on open data sets could also be covered by the GMG Data Access and Usage/Interoperability Working Group.

## 5. REFERENCES


Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal. C., Bengio, Y., & Shee, A. (2019). Towards standardization of data licenses: The Montreal data license. https://arxiv.org/pdf/1903.12262.pdf (arXiv:1903.12262 [cs.CY]), http://www.montrealdatalicense.com

Center for Open Science. (2020). TOP guidelines. https://www.cos.io/our-services/top-guidelines

Creative Commons. (2020). Attribution-NonCommercial-ShareAlike 4.0 International. https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Creative Commons Licenses. https://creativecommons.org/licenses/

Specific licenses cited: Creative Commons Attribution-Non-Commercial-ShareAlike (CC BY-NC-SA): https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

City of Toronto. (1998-2020). What is open data? https://www.toronto.ca/city-government/data-research-maps/open-data/what-is-open-data/

Dat Foundation Governance. (2020). Dat protocol. https://dat.foundation/about/history/

Gingras Little, K., McKinnon, S., Moreau-Verlaan., L., McDonald, A. (2017). Analysis of Strainbursts in the Sudbury Region and Numerical Modelling of Destress Blasting [Data Set]. Queen's University Dataverse. https://doi.org/10.5683/SP/4RFHBJ

Global Mining Guidelines Group. (2019). Foundations of AI: A framework for AI in mining. https://gmggroup.org/wp-content/uploads/2019/10/GMG_Foundations-of-AI-A-Framework-for-AI-in-Mining-2019-10-07_v01_r01.pdf

Google AI Blog. (2006). All Our N-gram are Belong to You. https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

Government of Canada. (2020). Open Data 101. https://open.canada.ca/en/open-data-principles

Harvard University Privacy Tools Project. (2021). Differential Privacy. https://privacytools.seas.harvard.edu/differential-privacy

International Organization for Standardization. (2018). ISO 31000 risk management. https://www.iso.org/iso-31000-risk-management.html

Lubarsky, B. (2017, April). Re-identification of "anonymized" data. Georgetown Law technology review. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/

Manca, G. (2020). "Tennessee-Eastman-Process" Alarm Management Dataset. IEEE Dataport. https://dx.doi.org/10.21227/326k-qr90

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., & DeHaven, A. C. (2016, October 5). Transparency and Openness Promotion (TOP) Guidelines. https://doi.org/10.1126/science.aab2374

Open Knowledge Foundation (2020). "What is Open Data." Open Data Handbook. http://opendatahandbook.org/guide/en/what-is-open-data/

Personal Data Protection Commission Singapore. (2018, January 25). Guide to basic data anonymisation techniques. Retrieved from https://www.pdpc.gov.sg/help-and-resources/2018/01/guide-to-basic-data-anonymisation-techniques

Sikorska, J., & Hodkiewicz, M. (2019, January 17). Flowchart of data set risk assessment tool, version 3. https://drat-process.com

Sikorska, J. Z., Fraser, R., Bradley, S., & Hodkiewicz, M. R. (2019). Data Risk Assessment Tool for industry-academic collaborations. https://drat-process.com

Werner, T., Taneja, L., Huijbregts, M., Northey, S.A., Schipper, A.M., & Mudd, G. (2019). Global-scale remote sensing of mine areas and analysis of factors explaining their extent. Global Environmental Change, 60. https://doi.org/10.1016/j.gloenvcha.2019.102007