Toward a Metadata Framework for Sharing Sensitive and Closed Data: An Analysis of Data Sharing Agreement Attributes

Sam Grabus[1], Jane Greenberg[2]

[1, 2]Metadata Research Center <MRC>, College of Computing & Informatics (CCI), Drexel University, Philadelphia, PA, USA
{sam.grabus, janeg}@drexel.edu

Abstract. Legal and policy-oriented restrictions often hamper if not inhibit well-intended efforts to share sensitive or restricted data. The research reported on in this paper is a part of a larger initiative to develop a prototype system for automatically generating data sharing agreements that address privacy, legal concerns, and other restrictions. A content analysis was conducted, examining a sample of 26 data sharing agreements. The results include 6 high level categories, 15 mid-level attributes, and over 90 lower-level specific attributes, a portion of which can help to expeditiously support the automatic development of data sharing agreements. The paper presents background information, research questions and methods, results, and a discussion. The conclusion summarizes our results and identifies next steps.

Keywords: Metadata · Closed data · Data sharing agreements · Restricted data · Privacy and data sharing

1. Introduction

Big data and the subsequent proliferation of data science and data analytics have bolstered data as a major avenue of research. These new disciplines offer the potential and promise to use data for gaining new insights and addressing societal and environmental problems locally and worldwide. As data has become a first-class information object, there is an increased drive to share data, and the open data movement has progressed as a significant development [1]. There is also a growing desire to share data that is sensitive or restricted, although this development has not progressed at the same pace, due to legal concerns and other restrictions [2, 3, 4].

The predicament with sharing closed or sensitive data is particularly detrimental when industry or government pursue an academic data sharing partnership and the plans fall apart, after considerable time and effort have been directed to the endeavor. This case is all too frequent, which results in a waste of resources, due to the legal fees and time commitments for those seeking the data sharing agreement. Another common scenario is that the data sharing negotiation simply takes too long, so that by the time the agreement is finally confirmed, the data being considered is no longer

desirable. Again, the financial, time, and human resources expended are seen as a drain on those involved.

Closed data sharing challenges were a major focus under a set of sessions entitled "War Stories" at the "Enabling Seamless Data Sharing in Industry and Academia" workshop, held at Drexel University (Fall 2016). The workshop also covered cases in which data sharing agreements were, in fact, successful, and presenters provided insight into lessons learned. An important outcome at this workshop was unanimous agreement regarding the need for a system that could automatically generate data sharing agreements. This outcome defines a key goal of "A Licensing Model and Ecosystem for Data Sharing" [5] initiative, being pursued as an NSF Spoke research project that is part of the North East Big Data Innovation Hub (NEBDIH) [6].

Many digital library and repository technologies provide options for secure, password-protected access, as well as controlling selected aspects of privacy, although these systems do not integrate fully with secure licensing models. The NSF Spoke research project (A Licensing Model and Ecosystem for Data Sharing) seeks to address this need by integrating data sharing and license development processes that extend beyond open creative commons standards. The Spoke research is being pursued through a collaboration involving Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory (MIT/CSAIL), Drexel University's Metadata Research Center, and Brown University's Computer Science department. We are working toward an environment that will leverage existing de-identification and anonymization developments, and integrate with the DataHub platform [7]. The research presented in this paper is part of this larger effort to advance approaches for sharing closed data and ensuring security throughout the data-lifecycle.

The goal of the research presented in this paper is to identify metadata categories and attributes that will automatically and expeditiously support the development of data sharing agreements in closed and restricted environments. Specifically, the paper reports on a content analysis examining a sample of 26 data sharing agreements, and the identification of metadata categories and attributes. The paper that follows includes background information on open data, closed data, and data sharing; presents our research questions and methods; reports the first phase results; and includes a contextual discussion. The conclusion notes key findings and next steps.

2. Data Access and sharing: A Simplified Continuum

Data access is directed by both contractual and technical factors. At the contractual level, there is a range of permissions, policies, legal considerations, personal and organizational preferences, and other factors that impact the data access rights. Rights, in this context, may cover permissions to view, use, reuse, repurpose, or distribute data. Metadata attributes, such as "rights management," can be assigned to data manually or automatically. When applied, rights management indicates data access status and use conditions. Data can be published or released under recognized licenses [8][9], which may be documented in the rights-oriented attribute, or appended to the data in another way. These conventions are primarily contractual, and inform technical aspects of system design. More significantly, such conventions have

helped to advance data access and sharing across a broad range of communities. Progress aside, it is important to realize that the amount of data published and released represents only a fraction of the vast quantities of data generated daily, including valuable data that owners may like to share, but can't, due to a myriad of restrictions. To understand the complexities of data access, both contractual and technical, it is helpful to first review the status of data access — specifically, what is meant by open and closed data.

2.1 Open Data

Open data is "data that anyone can access, use or share" [10]. More precisely, the Open Data Handbook explains that "open data is data that can be freely used, reused and redistributed by anyone — subject only, at most, to the requirement to attribute and sharealike" [11]. This resource elaborates on this definition by explaining important qualities, such as the practice that open data must be available and accessible, reusable, and redistributable, without restriction [11]. The absence of restriction surrounding open data extends to any endeavor, including commercialization. As noted, there are a range of licenses that data producers or data hosts append to data, indicating open access. The Creative Common Zero [12] is perhaps the most commonly known set of licenses.

While open data is increasingly lauded as beneficial and necessary to advance solutions for societal problems, the bioinformatics data community is arguably becoming more closed, due to the increasingly private nature of the data itself [13]. The sensitive nature of bioinformatics data falls into the category of closed data, given the need to protect individuals from potential exploitation and the harm that could result from research subject re-identification.

2.2 Closed Data

Closed data is data that often contains private or sensitive information. Closed data extends across a wide range of entities, topics, and environment. Examples of closed data include personal, institutional, or industry data identifying financial resources (e.g., sums, transactions, account numbers), personal information relating to health and well-being, or status (e.g., married, single, divorced). Data may be designated as closed, or regulated by controlled access, due to legal restrictions or organizational policies protecting current or predicted value. More specifically, data access is often restricted because of a known or perceived competitive advantage, and the associated risks with making it public, including misuse, if the data falls into the hands of the wrong person, office, or organization.

The last several years, there have been a series of data breaches exposing closed data, causing mayhem and outcomes that been ruinous to individuals. A case in point, the Ashley Madison hack [14], in which a torrent was used to publish user profiles, transactions, credit card data, and a wide range of other sensitive data, including associated metadata. Ashley Madison's core mission is to arrange extramarital relations between married individuals, and the breach exposed the private lives of thousands of people, damaging the reputation and status of people who used their work email addresses to use the website.

There are also cases that raise concern about potential exploitation, in which private data is shared between organizations without the consent of the data subject. For example, UK's National Health Service allegedly violated data protection laws by sharing patient health data with Google's DeepMind [15]. Examples of medical data sharing frequently raise public concern, particularly with an increase of more than 20% in cases of medical identity theft between 2013 and 2014 [16].

These examples of sharing closed data demonstrate potentially harmful impacts, but this does not mean that closed data should not be shared, considering the immense societal advantages and innovation potential from new insights.

2.3 Data Access Status: Fuzzy Boundaries

To be clear, closed data is accessible to individuals or organizations who have the appropriate permissions. In other words, access and use are conducted in a tightly controlled, secure environment. Additionally, technical system features are implemented to support and maintain the controlled environment, albeit sometimes hackable, as demonstrated in the above and other cases. A common scenario is that data may initially be closed, marked by an embargo period, and then later be released as open. This option is one seen in the Dryad data repository [17], for published data that underlies published research. Additionally, data that is designated open may have components that need to remain closed, such as personally identifying information (PII). This is often the case when personal health data is made available for medical research. For example, the progression of a disease is important for research, but individuals who have certain disorders may want to protect their identity [18]. Approaches to sharing sensitive data include anonymization and de-identifying personal information [19], allowing some, but not all, of the data to be shared. This example, and others alluded to here, demonstrate the fuzzy boundary between open and closed data, and impact the increasing trend toward data sharing.

3. Data Sharing Practice and Needs

For the last decade and a half, extensive energy has been directed toward cultivating and sustaining open research and the sharing of creative output. Many of the ideas stem from the open source community, and the notion of wisdom of the crowds [20]. Simply put, the more people contributing to R&D, the better the end results. Additionally, open sharing environments keep people from reinventing the wheel. Specific to science, there have been considerable efforts to build cyberinfrastructure for sharing scientific and scholarly outputs [21], with data becoming a key focus.

Another significant motivator of data sharing has been the data deluge—that is, the sheer amount data generated through scientific research and other endeavors, and the unprecedented capacity to support data-driven science [22]. These ideas stem from Jim Gray's notion of the Fourth Paradigm enabling highly efficient access to data and analysis tools [23].

Data sharing in the closed environment has benefitted open data developments, where tools and technologies developed can be used and made secure. Even so, the rate of progress is hampered by legal and logistical challenges. Although useful infrastructure can help facilitate secure data sharing, a range of regulations and policies (e.g., HIPAA, institutional policies) complicate closed data sharing, impeding or otherwise delaying the process.

Research needs to be directed toward developing solutions to address these data sharing challenges in the closed environment. Metadata has been identified as part of the solution to addressing these challenges [24]. To move forward with metadata solutions, more analysis is needed to determine researcher data sharing needs. The research that follows takes initial steps toward addressing this need.

4. Research Questions

The overriding goal of this research is to advance metadata practices that can facilitate data sharing efforts in environments where the data is not necessarily open. The specific goals were to identify metadata categories and attributes that will, more expeditiously, support the development of data sharing agreements in these environments. The research was guided by two questions:

1. What high-level metadata categories are found in data sharing agreements?
2. What elements and attributes are found across data sharing licenses with sensitive information; what are the most common?

The next section reports methods and the steps taken to investigate our questions.

5. Methods and Procedures

To address the questions posited above, we conducted a content analysis following the general eight steps outlined by Zhang and Wildemuth [25]. Our research protocol involved the following steps:

1. *Data collection*: A sample of 26 data sharing agreements from industry, academia, and government was obtained. These agreements were collected via a solicitation, following the NE Big Data Hub's Data Sharing Workshop, and other communications through the NSF Big Data Innovation Hub Program, and in connection within the Research Data Alliance. The data sharing agreements were acquired through a two-step process. First, a number of licenses were found through online searching, which typically yielded templates containing useful language, along with sections to be completed by the parties involved in the data sharing. Second, a call was distributed to the diverse community that participated in the workshop, with members from industry, government, and academia. Achieving a representative sample across all sectors was a challenge; and our sample emphasized academic partnerships. Additionally, agreement dates were

difficult to confirm. Despite these issues, the convenient sample of 26 agreements included representation from multiple sectors, including industry, and was deemed sufficient for our research.

2. *Content analysis:* The language of these licenses was first examined for overall clarity and to confirm data sharing support; then, a focused examination was pursued. The language was parsed for higher-level general categories, as well as mid and lower-level attributes, which are the detailed license specifications for data handling. The full collection of categories and attributes make up the sample for the study reported on in this paper.
3. *Language clustering:* The high-level general categories and lower level attributes were organized on a spreadsheet in a hierarchical arrangement.
4. *Metadata labeling*: The language of the categories and attributes was refined. These results will be used to help build the infrastructure that will assist with the minting of data sharing agreements.

6. Data Analysis

The data analysis was carried out to answer our two questions, looking at high-level categories for sharing data, as well as common elements and attributes among the data sharing agreements. In answer to question 1, we found that the agreement attributes fit into 6 high-level broad categories, listed in **Figure 1** below. This categorization is the first iteration of refinement, and as such, we recognize that there may be some dependencies or overlap:

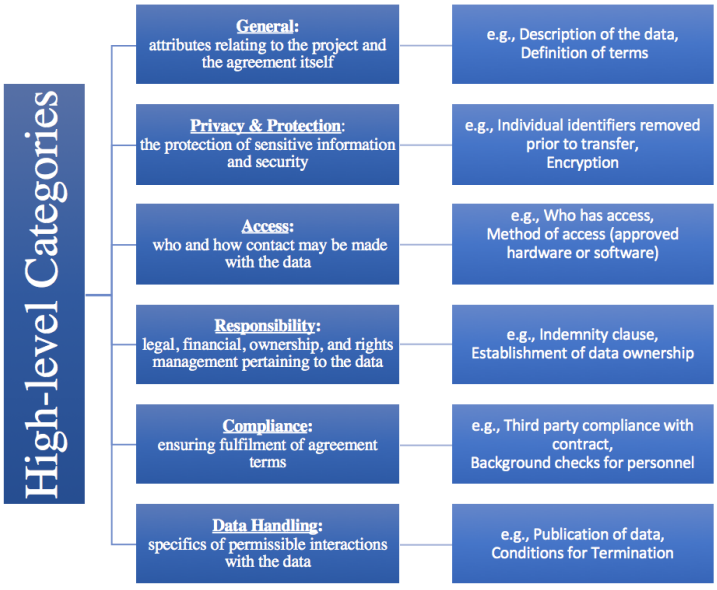

Fig. 1. High-level Categories

In addressing question 2, we discovered 15 common mid-level elements (e.g., Legal, Termination), and over 90 lower-level attributes that address the specific data sharing needs (e.g., Who has access to proprietary information, Merging data with other sets).

Following the high-level categories, Table 1 (below) presents detailed results for "Privacy & Protection," and Table 2 presents details of "Data Handling." The common elements were gleaned from the 26 data sharing licenses across the full sample.

Table 1. Privacy & Protection Attributes

<table>
<tr><th colspan="3">Privacy & Protection</th></tr>
<tr><th colspan="3">Sensitive Information</th></tr>
<tr><th>Regulations</th><th>Preparing data</th><th>Access</th></tr>
<tr><td>• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)
• Compliance with federal/state/international data protection laws and regulations</td><td>• Identification of confidential/special categories of information (e.g., pii, proprietary)
• Individual identifiers removed/anonymized prior to transfer</td><td>• Who has access to pii/confidential data
• Who has access to proprietary information</td></tr>
<tr><th>Privacy</th><th>Avoiding re-identification</th><th>Exceptions</th></tr>
<tr><td>• Anonymization of data
• Confidentiality and safeguarding of PII/sensitive data
• Removal/nondisclosure of company/personnel identification in materials and publications
• No contact with data subjects</td><td>• No direct/indirect re-identification
• Statistical cell size (how many people, in aggregated form, can be released in groups)
• Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify</td><td>• Exceptions to confidentiality
• Conditions of proprietary information disclosure
• Conditions of pii disclosure (who, what, and for what purpose?)
• Limitations on obligations if data becomes public
• Limitations on obligations if data is already known prior to agreement
• Limitations on obligations if data given by 3rd party without restriction</td></tr>
<tr><th colspan="3">Security</th></tr>
<tr><td colspan="3">• Sharing non-confidential data
• Password protection/authentication of files
• Encryption
• Security training for involved personnel
• Establishing infrastructure to safeguard confidential data</td></tr>
</table>

Table 2. Data Handling Attributes

Data Handling		
Use		***Physical***
• Each data field/elements to be accessed • Use of data: only for project-specific/research, or analytical use • Documenting all projects using the data	• Modification of data • Compliance with data updates (changes, removal, corrections) • Sharing data	• Copy/reproduction of data • Storage of data • Transfer of data (e.g., allowed methods)
Results		***Personal Gain***
• Presentation of data • Publication of data (e.g., prior approval needed or right to publically disclose publication)	• Results/reports and associated documents (e.g., must be provided copies) • Right to remove/delete confidential data from proposed publications	• Sale of/profit from data (e.g., noncommercial use only) • Licensing of data • No reverse engineering
Termination		
• Conditions for termination • Destruction or return of data after agreement • 3rd party destruction or return of dataset • Confirmation of data destruction	• Data retained or used for period of time after termination • Which rights and obligations remain in effect after termination	

In addition to the identification of categories and attributes, we tallied the most common attributes. These percentages reported below are based on the 90+ lower-level attributes discovered among the full sample of 26 data sharing agreements:

- Most common attributes pertaining specifically to protecting sensitive information were adherence to federal/state/international data protection laws and regulations, as well as explicitly prohibiting direct or indirect re-identification of data subjects, found in over half the sample of agreements (16 of 26, 61.5%).
- Agreements involving industry had the highest instances of the data ownership attributes.
- The most common attribute found among general data handling practices, across the sample, was the return or destruction of data after the agreement ends (88.4%).
- Over half (61.5%) of all agreements specify which rights and obligations will remain in effect after termination.

- Over half (65.3%) of all agreements specify that the data should only be used for the specific research and analytic uses agreed upon through the license.
- Almost half (46.1%) of data sharing agreements include an indemnity clause, specifying that the other party will be held responsible for damages involved with the data usage.
- The most common privacy laws applicable to the data sharing agreements are HIPAA (Health Insurance Portability and Accountability Act) and FERPA (The Family Educational Rights and Privacy Act).
- Only 15.3% of the agreements specify whether the data can be merged with other sets (in aggregated form or otherwise).
- Only 4 (15.3%) of the agreements required source acknowledgement for use of the data.

7. Discussion

The results presented above are important to the research being conducted as part of the "A Licensing Model and Ecosystem for Data Sharing" project. Our results can connect with efforts underway with the Dataverse system, where a color-coded tag system has been developed to designate degree of data access [26]. One end of the spectrum is a light blue, to designate open access, and at the other end is a deep crimson red, indicating "maximally restricted," with access requiring a "Two-factor authentication, Approval, Signed DUA." [26]

As expected, all of the licenses in the sample contained attributes regarding the privacy and protection of sensitive data, whether it is personally identifiable information (PII) or proprietary information. The level of specificity for these protections varied drastically, but the most common attributes pertaining specifically to protecting sensitive information were adherence to federal/state/international data protection laws and regulations (61.5%), as well as explicitly prohibiting direct or indirect re-identification of data subjects (61.5%). This makes sense, considering that presumably all stakeholders would want to avoid the potential legal repercussions of violating privacy laws, particularly in the event that data subjects are re-identified. The results here will help in developing our framework to support privacy, and ensure data integrity. Advances by Barth, et al [27], in which they developed a framework to incorporate legislation found in HIPPA, COPPA, and GBLA, and support reasoning about the transmission of personal information, may further inform our research in this area.

Although many of the licenses examined were from academic institutions, they all identified attributes related to privacy. Academia, industry, and government averaged an almost identical number of attribute occurrences related to privacy and security concerns (Academia had an average of 5.45 attributes in this group, Industry: 5.71, Government: 5). Close to a third (27%) of the agreement sample specifically involved industry, with 15.4% involving government. Although we are continuing to build our sample base, the sample used was sufficient for this initial study, and is already

informing our prototype system design. Specifically, research team members have started to develop a prototype that will interconnect with DataHub, and the results presented above have helped us to determine which facets can more easily be automated. This determination has helped in more clearly identifying areas of greater priority for automation and implementation (e.g., automating the removal of PII categories, as well as aspects of agreement access and termination). The results have also been important for understanding where our system must, instead, support opportunity for textual phrases, either accessible via a dropdown menu, or through the flexibility to generate desired statements for the parties pursuing the agreement. This mixed approach is common across many data systems, from library serials control [28] to data repositories, such as the Morpho system [29], for ecological data. Automated controlled attributes, as well as open text options, will be features in our system, supporting data sharing agreements, permissions, and licensing.

8. Conclusion

Data sharing of sensitive and private information introduces a set of challenges, and requires the development of most robust agreements, far different than with open data, where data sharing is established with an open license. The research presented in this paper is part of a larger effort to expedite the process of developing data sharing agreements. The research focused on metadata aspects that can help with the automatic generation of agreements to support data sharing in these more restrictive environments. A content analysis examining a sample of 26 data sharing agreements was conducted. Key findings were:

- 6 high-level categories
- 15 mid-level attributes, and over 90 lower-level specific attributes, which can help to inform the development of automatically-generated data sharing agreements
- Observations that showed a prevalence of particular privacy and data handling attributes over others

The results have already proved useful in taking steps to help develop a prototype system. Research team members have already taken steps to automate the removal of personal identification information, specify controlled access for particular researchers to specific tables/cells of data, and execute termination of access to some or all of the dataset at the end of the agreement. In addition to aiding the development of the data sharing platform, this paper also shares our initial methods, which may help other researchers pursuing similar work. While the results presented in this research have helped with our next steps, we recognize the sample limitations, and intend to address this challenge by developing a process that will allow to obtain a more representative sample of data sharing agreements.

Next steps include developing and analyzing a larger sample of data sharing agreements, to confirm our initial findings and gain a more comprehensive understanding of stakeholder data sharing requirements, particularly when privacy, sensitive information, and other restrictions are involved. Other next steps include minting key phrases that associate with our metadata categories — specifically, the mid- and lower-level attributes — that would be inserted into data sharing

agreements. Longer-term goals will include user review of the metadata attributes, and testing the prototype system being developed by other key members.

Data may hold the solutions to addressing many of society's grand challenges. To this end, researchers need to pursue steps to improve data sharing across all environments, including those with sensitive/private/restricted information. The research presented here is a strong effort toward addressing the current data sharing challenges, and our next steps forward will further contribute to metadata and data sharing research initiatives in the future.

Acknowledgements
We acknowledge the support of the National Science Foundation/IIS/BD Spokes/Award #1636788; and thank Sam Maddden (MIT), Carsten Binnig (TU Darmstadt), and Tim Kraska (Brown University), and other individuals who provided us with data sharing agreements.

References

1. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268.
2. McGuire, A. L., Oliver, J. M., Slashinski, M. J., Graves, J. L., Wang, T., Kelly, P. A., ... & Treadwell-Deering, D. (2011). To share or not to share: a randomized trial of consent for data sharing in genome research. *Genetics in medicine: official journal of the American College of Medical Genetics, 13*(11), 948-955.
3. Pencarrick Hertzman, C., Meagher, N., & McGrail, K. M. (2012). Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. *Journal of the American Medical Informatics Association, 20*(1), 25-28.
4. Gleason, C. J., & Hamdan, A. N. (2017). Crossing the (watershed) divide: Satellite data and the changing politics of international river basins. *The Geographical Journal, 183*(1), 2-15.
5. Metadata Research Center (2017). *A Licensing Model and Ecosystem for Data Sharing*. Retrieved from `http://cci.drexel.edu/mrc/projects/a-licensing-model-and-ecosystem-for-data-sharing/`
6. Northeast Big Data Innovation Hub. (2017). Retrieved from `http://nebigdatahub.org/`
7. Datahub (2016). *What is datahub?* Retrieved from `https://datahub.csail.mit.edu/www/`
8. Creative Commons (2017). *Licensing types.* Retrieved from: `https://creativecommons.org/share-your-work/licensing-types-examples/`
9. The National Archives (n.d.). *Open government license for public sector information.* Retrieved July 1, 2017, from: `http://nationalarchives.gov.uk/doc/open-government-licence/version/3/`
10. Open Data Institute (n.d.). *What is open data?* Retrieved July 1, 2017, from `https://theodi.org/what-is-open-data`
11. Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J., & Zijlstra, T., et al. (2009). *Open data handbook*. Retrieved June 15, 2017, from `http://opendatahandbook.org`

12. CC0. (n.d.). Retrieved July 1, 2017, from https://creativecommons.org/share-your-work/public-domain/cc0/
13. Greenbaum, D., Sboner, A., Mu, X., & Gerstein, M. (2011). Genomics and privacy: Implications of the new reality of closed data for the field. Plos Computational Biology, 7(12), e1002278. doi:10.1371/journal.pcbi.1002278
14. Segall, L. (2017). Ashley Madison: Life after the hack. *CNN Tech*. Retrieved from http://money.cnn.com/mostly-human/click-swipe-cheat/
15. Kwon, D. (2017). Google's DeepMind, UK's NHS criticized for sharing data. *The Scientist*. Retrieved from http://www.the-scientist.com/?articles.view/articleNo/49812/title/Google-s-DeepMind--UK-s-NHS-Criticized-for-Sharing-Data/
16. Kaplan, B. (2016). How should health data be used? *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees*, 25(2), 312.
17. Dryad Digital Repository. (2017). Retrieved from http://datadryad.org/
18. Liu, X., Li, X., Motiwalla, L., Li, W., Zheng, H., & Franklin, P.D. (2016). Preserving Patient Privacy When Sharing Same-Disease Data. *Journal of Data and Information Quality 7* (4). https://doi.org/10.1145/2956554
19. El Emam, K., et al. (2012). De-identification methods for open health data: The case of the heritage health prize claims dataset. *Journal of Medical Internet Research, 14*(1), e33.
20. Raymond, E. (1999). The cathedral and the bazaar. *Philosophy & Technology, 12*(3), 23. Retrieved from https://monoskop.org/File:Raymond_Eric_S_The_Cathedral_and_the_Bazaar_rev_ed.pdf
21. Atkins, Daniel E. (Daniel Ewell), & National Science Foundation (U.S.). Blue-Ribbon Advisory Panel on Cyberinfrastructure. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure*
22. Hey, T. and Trefethen, A. (2003) The Data Deluge: An e-Science Perspective, in *Grid Computing: Making the Global Infrastructure a Reality* (eds F. Berman, G. Fox and T. Hey), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/0470867167.ch36
23. Hey, T., Tansley, S., Tolle, K. (2009) *The fourth paradigm: Data-intensive scientific research*. Microsoft Research. Redmond, WA.
24. Greenberg, J., Grabus, S., Hudson, F., Kraska, T., Madden, S., & Bastón, R. (2016). *The Northeast Big Data Hub: "Enabling Seamless Data Sharing in Industry and Academia" Workshop.* Philadelphia, PA: The Northeast Big Data Innovation Hub. https://doi.org/10.17918/D8159V
25. Zhang, Y., & Wildemuth, B. M. (2005). Qualitative analysis of content, Applications of social research methods to questions in information and library science. 2009. *Google Scholar*, 308-19.
26. Crosas, M. (2016). The DataTags system: Sharing sensitive data with confidence. *RDA 8th Plenary*, Denver Colorado, Sept. 16, 2016. Retrieved from https://scholar.harvard.edu/mercecrosas/presentations/datatags-system-sharing-sensitive-data-confidence
27. Barth, A., Datta, A., Mitchell, J. C., & Nissenbaum, H. (2006, May). Privacy and contextual integrity: Framework and applications. In *Security and Privacy*, *2006 IEEE Symposium* on (pp. 15-pp). IEEE.
28. Blake, K., & Collins, M. (2010). Controlling chaos: management of electronic journal holdings in an academic library environment. *Serials Review, 36*(4), 242-250.
29. Higgins, D., Berkley, C., & Jones, M. B. (2002). Managing heterogeneous ecological data using morpho. *Paper presented at the 14th International Conference on Scientific and Statistical Database Management,* 69-76. doi:10.1109/SSDM.2002.1029707