Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer

Schembera, Björn; Duran, Juan Manuel

DOI
10.1007/s13347-019-00346-x

Publication date
2019

Document Version
Final published version

Published in
Philosophy & Technology

Citation (APA)
Schembera, B., & Duran, J. M. (2019). Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer. *Philosophy & Technology*, *33*(1), 93-115. https://doi.org/10.1007/s13347-019-00346-x

Philosophy & Technology
https://doi.org/10.1007/s13347-019-00346-x

RESEARCH ARTICLE

Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer

Björn Schembera[1] **· Juan M. Durán**[2]

Received: 25 June 2018 / Accepted: 1 March 2019 / Published online: 13 March 2019

Abstract

Many studies in big data focus on the uses of data available to researchers, leaving without treatment data that is on the servers but of which researchers are unaware. We call this *dark data*, and in this article, we present and discuss it in the context of high-performance computing (HPC) facilities. To this end, we provide statistics of a major HPC facility in Europe, the High-Performance Computing Center Stuttgart (HLRS). We also propose a new position tailor-made for coping with dark data and general data management. We call it the *scientific data officer* (SDO) and we distinguish it from other standard positions in HPC facilities such as chief data officers, system administrators, and security officers. In order to understand the role of the SDO in HPC facilities, we discuss two kinds of responsibilities, namely, technical responsibilities and ethical responsibilities. While the former are intended to characterize the position, the latter raise concerns—and proposes solutions—to the control and authority that the SDO would acquire.

Keywords Research data management · High-performance computing · Dark data · Big data · Computer simulations · Scientific data officer · Data curation

Both authors contributed equally to all sections of the paper.

✉ Juan M. Durán
j.m.duran@tudelft.nl

Björn Schembera
schembera@hlrs.de

[1] High-Performance Computing Center Stuttgart, University of Stuttgart, Nobelstr. 19, 70569 Stuttgart, Germany

[2] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, Netherlands

1 Introduction

Studies in big data across disciplines are chiefly interested in the analysis and uses of actual data. Social networking, business, and governments are typically, although not exclusively, interested in understanding the implications of big data in the social behavior, business practices, and the role of the government presence in the citizen's life (Mayer-Schönberger and Cukier 2013; Glass and Callahan 2014). Big data in scientific and engineering research offers large amounts of data at the researcher's disposal, to the extent that new avenues of investigation and discovery can be transient. Excellent work on the uses and implications of big data in scientific research is evidenced in neuroscience (Choudhury et al. 2014), biology (Leonelli 2013, 2014), and astronomy (Edwards and Gaber 2014), just to mention a few scientific disciplines. The universe of big data, of which we have only mentioned a few cases, is chiefly interested in "tangible data," that is, data that is knowledgeably available to the user (e.g., researchers, public servants, businessman and businesswoman). Opposing "tangible data" is "dark data," understood as data that is stored in the servers available for use, but of whose existence users are unaware. The specialized literature has been deeply involved in understanding "tangible data," but it has largely neglected "dark data." This article calls attention to the other side of data, the one that we call dark for being potentially rich sources of reliable information, but which is forgotten on storage servers away from any possibility of usage. Since the kind, amount, and uses of data in scientific and engineering research are rather different from social networking, business, and governmental institutions, we shall focus on dark data in high-performance computing (HPC) facilities.

Under ideal conditions of scientific practice, standard data management workflows in HPC facilities indicate that, in order to keep clean records of the data produced, such data must be "labeled" correctly. This means that specific *metadata* about the data is tagged onto the data with the purpose of identification and categorization. Examples of metadata include the date and time stamps of when the data was created and modified, a string containing the full directory and ownership, and descriptive information on the content of the data. Metadata, then, plays the fundamental role of structuring, informing, and identifying data by means of relevant information about them.

Now, when such conditions of management workflow are not followed, and admittedly this is the case for many HPC facilities, then data becomes dark, invisible, and undetectable by the researchers. To us, this represents a significant problem for data management in HPC facilities that needs to be addressed. The presence of dark data is particularly pressing when these facilities are not part of the private sector, but rather are affiliated with public universities. Our example is the High-Performance Computing Center Stuttgart (HLRS), located at the heart of the University of Stuttgart, one of the main supercomputing centers in Europe and the world, and one of the largest of its kind in Germany.[1] The distinctive characteristic of an HPC facility such as the HLRS

[1]The HLRS is actively involved in reducing dark data from their servers. Based on this study, measures such as a check of the data inventory have been performed and dark data was identified. The HLRS is currently discussing concrete workflows aiming at diminishing dark data. For a ranking on HPC centers around the world, see https://www.top500.org/.

is that the data produced has importance in scientific and engineering processes, and therefore the cost of production, as well as the social and environmental benefits, are significantly different from other forms of big data. To put it into perspective, consider the costs of training an engineer and add to it the costs of producing scientific data; now compare these costs with collecting data from companies such as Facebook or Google. In the latter case, algorithms are typically involved in harvesting data from the user's access to their profiles, as opposed to actual researchers producing data. Additionally, and since we are interested in the case of public institutions, the benefits and costs of producing, processing, storing, and generally managing data are shared by the entire population. In contexts where big data is used for establishing the shopping preferences of users, profits and gains—whether financial or societal—typically remain with the company that invested in the computational systems and the necessary infrastructure.

Taking the notion of dark data as "not carefully labeled data" as the basis for our analysis, we propose to build a new definition of dark data exclusively tailored to the current practice in HPC facilities. Additionally, this article elaborates on the *scientific data officer* (SDO), understood as a new position for HPC facilities with the responsibility of handling dark data. As we will show, this position differs from standard ones found in HPC facilities (e.g., system administrator, chief data officer), turning the SDO into a major role for the future of data management.

The article is then structured as follows. Section 2 presents the problem of dark data as discussed in the current literature. Although we agree with most definitions, we will suggest our own approach specifically tailored to dark data in HPC facilities. Section 3 presents current statistics on dark data at the HLRS. We use this section to exemplify the seriousness of the problem. Next, Section 4 attempts to introduce a new role for managing data, namely, the SDO. This section deals, therefore, not only with a proper characterization of the SDO but also with distinguishing the SDO from other data management positions already existing. Finally, we use Section 5 to suggest how dark data could be handled and returned to the disposal of the scientific community, complying with basic standards of the Open Data movement.

2 The Problem and Notion of Dark Data

Contemporary science and engineering research are producing extremely large amounts of data in the order of PetaBytes that, as standard scientific practice dictates, must be stored in designated storage facilities. HPC facilities are at the heart of these new forms of scientific and engineering research, usually present by means of complex computer simulations as well as the work driven by big research data. It then comes as no surprise that HPC facilities are a main player in the management of scientific and engineering data. Now, a major challenge that HPC facilities are facing today has its roots in the curation, storage, and general maintenance of large amounts of data produced by researchers. While much of the attention is devoted to the early stages of the data life cycle, such as the production and short-term management of the data (Leonelli 2013, 2014), much less attention is being put on long-term storage handling and maintenance of data.

It is frequently the case that researchers who wish to share their data need to make sure that both the data and the metadata they are submitting are compatible with existing standards. Incompatible data might be as good as noise, for it cannot be processed for further use and is therefore useless for research. This means that researchers must find time in their already busy agenda not only for standardizing their data and metadata but also for acquiring updated knowledge on the standards used by a given repository (Leonelli 2014, 4). Unfortunately, academia and other institutions are not always receptive to their researchers spending qualitative time on such endeavors. As a result, these unacknowledged—but implicitly required—efforts give researchers very few incentives to share, format, and standardize their data and metadata for further use, with the subsequent loss of professional collaboration and research efforts. It then takes very little for good, reusable scientific data to be forgotten in large storage servers despite the costs of producing, gathering, curating, and storing such data. We call such data *dark data* as a metaphor for their lack of visibility in much contemporary scientific and engineering research.

Dark data can be understood better as data stored in designated storage servers but which is no longer in use—or nearly not in use—by any user, whether this means the original user that created the data, or new users who might be interested in reusing the data. It is "dark" because this data becomes invisible by means of not being easily found or accessible by any user or administrator. Instead, it remains unused in the storage servers allocating resources that could be redirected to more useful ends.

A definition of dark data—especially in the context of HPC facilities—is virtually non-existent.[2] A first characterization of dark data in the context of scientific activities comes from the work of Bryan Heidorn, who defines it as data that is "not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost"[3] (Heidorn 2008, Abstract). Following this definition, dark data arises only in the so-called long tail of science, which is defined as a large number of small research projects that have neither time nor funding to accomplish the task of proper data management. Such large amounts of small projects would always, according to Heidorn, enforce the presence of dark data since it is "the type of data that exists only in the bottom left-hand desk drawer of scientists on some media that is quickly aging and soon will be unreadable by commonly available devices" (Heidorn 2008, 281).

While Heidorn's definition is sound and intuitively appealing, it is largely unusable for studies on data in HPC facilities. There are several reasons for this. First, Heidorn takes dark data to be related to *portable storage*. To his mind, data becomes

[2]The notion of dark data can be found in business contexts, although it is not fully developed. One approximation takes them as "as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing)" https://www.gartner.com/it-glossary/dark-data. For more characterizations of dark data in business, see Austin (2014), Shahzad (2017), Dennies (2015), and Babcock (2015). Other definitions can be found in the context of big government, such as Brantley (2015). Finally, there are studies on the notion of data in general (e.g., Barberousse and Marion (2013), Humphreys (2013), and Leonelli (2015)).

[3]Heidorn speaks of data indexing as data labeling with additional information to describe the data. This must not be confused with a database index.

dark when a portable disk gets lost in the drawers of some researcher. This particularity makes dark data to be of a special kind, one where the physical deterioration could easily be the cause of data becoming inaccessible regardless of the fact that researchers know about their existence. In HPC research, data becomes dark because researchers are unaware of their existence, despite their accessibility. Furthermore, Heidorn is only concerned with the increase of dark data insofar as it entails the growth of physical space in the researcher's office. A central worry about dark data in HPC facilities is that this growth is physically invisible and therefore it tends to be increased as new storage is added to the servers—a fairly common practice in HPC facilities.

In a recent paper, Heidorn returns to discussing dark data by showing how small research projects in astronomy produce dark data due to insufficient funding for research data management tasks (Heidorn et al. 2018). Admittedly, in this article, Heidorn et al. acknowledge dark data to also be found in digital format, and thus closer to our interests. However, Heidorn et al. do not seem to have special concerns for HPC research, where large amounts of data are produced from a small number of projects. Rather, their interests still lie on the so-called long tail of science, defined as a large number of projects producing small quantities of data. This is a fundamental distinction that separates our notion of dark data from that of Heidorn et al., specifically because the maintenance of data can be better handled (e.g., by standardization protocols held within the HPC facility) and it is centralized under the scientific data officer.

There is a third issue with Heidorn's conceptualization of dark data stemming from the distinction made between the "long tail of science" and the "head data" (Heidorn 2008, 285). Briefly, Heidorn organizes scientific projects along an axis from large to small projects, the latter engaging only a handful of researchers while the former encompasses dozens or even more scientists (Heidorn 2008, 281). The long tail of science differs from the head data in that, in the long tail of science, "data is more difficult to find and less frequently reused or preserved" (Heidorn 2008, 281). Instead, head data is claimed to be homogeneous, maintained, central curation tasks exist, and open access of this data to all authorized researchers is well established. Thus understood, only in the long tail of science can one find large amount of research data, not in the head data. In HPC facilities, this is clearly not the case. As mentioned, a typical situation is having a small number of multimillion euro projects with large amounts of data produced and to be produced. Such contexts thrive from a few characteristics found in the head data: the amount of curation could be reduced and, to a certain extent automatized since curators do not have to familiarize themselves with too many projects, the metadata protocols can be standardized within the HPC facility, fostering exchange of data for similar projects, and data and metadata can be within the orbit of a centralized authority (i.e., the SDO).

More generally, none of Heidorn's considerations on dark data seem to apply to HPC research (Schembera and Bönisch 2017). Data maintenance in HPC facilities is mostly handled by directory structures and filenames, rather than well-structured metadata standards. This makes data non-homogeneous in HPC facilities, a constraint not under consideration by Heidorn. As for the preservation of data, we mentioned earlier that to Heidorn's mind, data is related to portable data, and therefore it is most

likely that this concern relates to data enduring in unreliable media. This is less of a concern for data stored in HPC facilities due to technologies such as RAID and tape storage. Having said this, preservation might also mean that data survives regular purges of the servers by being put into a tape archive system. As we will argue later, there are well-established workflows that lead to storing data for long periods of time, removing the need to rely on physical media. Such workflows, as it turns out, also encourage the existence of dark data on the servers. Finally, we shall not forget that the kind of scientific and engineering activities and projects carried out in HPC facilities require special hardware and software, meaning that such research is virtually impossible outside these facilities. Dark data, then, are not solely produced by a careless researcher (i.e., one that forgets the hard disks in the bottom left-hand desk drawer), but principally by the research facility that chooses to give their researchers the responsibility of curating, managing, and storing their own data.

Another definition of dark data comes from the work of Thomas Goetz, who takes the notion to refer to data that has not "made the cut" for publication. Such data counts as dark because it is a source of knowledge that does not reach the scientific and engineering community. As Goetz puts it, "your dead end may be another scientist's missing link, the elusive chunk of data they needed" (Goetz 2007). Although Goetz's characterization of dark data depends on the large production of data, it is heavily reliant on the publishing industry, and to a certain extent academia as well. Just as Heidorn's definition, then, it is largely inapplicable to our study.

A common ground between the two notions discussed, and upon which we agree, is that data is "dark" because it becomes *invisible* somehow (i.e., not carefully indexed, as Heidorn claims, and not publishable, as Goetz believes). However, as we argued, our position extends Heidorn's and Goetz's in several ways. For starters, our notion of dark data is exclusive to HPC facilities. This means that our notion depends on large amounts of data that are not visible to a special community of researchers (i.e., those with access to supercomputing). Also, our notion takes dark data to be data whose proper metadata is missing, and in this respect, it is part of a technical feature of data, rather than considerations about researchers indexing their data (e.g., in a notebook) or having the data available via publications. The sources of mislabeling data vary from data that never was labeled with proper metadata, to non-standard labeling by the researchers, including changes in the file system that ends up hiding data that receives no maintenance. Another characteristic not considered by Heidorn and Goetz is that a major source of dark data is when researchers leave the HPC facility. In many cases, researchers leave behind data stored on the servers that, despite being indexed, become dark because the institution does not take the necessary measures to keep the data properly stored on their servers. These and other issues will be the subject of the following sections.[4]

[4]It is worth pointing out the parallels in HPC facilities to other laboratory-oriented science where data management is left to graduate students who move on at a high rate, taking the metadata with them (Darch and Sands 2015). Data "in" a laboratory is now frequently on a cloud service, somewhere outside of the physical bounds of the laboratory and perhaps in HPC facilities. We thank an anonymous reviewer for this clarification.

3 The Statistics of Dark Data at the HLRS

We now illustrate a case of dark data using current statistics obtained from the High-Performance Computing Center Stuttgart (HLRS), University of Stuttgart. First, two main sources of dark data (DD) are distinguished among the universe of users at the HLRS: users that are *de-registered*, and users that are *inactive* (see Fig. 1). Before fully discussing the statistics, let us first clarify the nature of each user and why they are considered a source of dark data.

As its name suggests, *de-registered* users are those researchers that have been deleted from the user administration database. Typically, this is done either by the researcher herself or by the system administrator. Thus understood, the data remaining on the servers from a de-registered user is considered potentially dark data, for statistics show that no other user (e.g., a researcher working on the same project as the de-registered user) will retrieve such data for subsequent use. Let it be noted that many institutional recommendations provided by funding agencies indicate that data needs to be available on the servers for a given amount of time—to be discussed shortly—for the purpose of subsequent use or in case they are requested by the owner or user again (DFG 2013, 74. Recommendation 7). In this respect, it is being assumed that all of the user's data is stored in the appropriate format, following standard procedures and protocols for future retrieval.

Having said this, it is important to emphasize again that a major source of dark data is when data is not properly labeled. Thus, if the de-registered user made use of a personal format for naming her own data, one which does not match the standards used or other user's format, then the data as it stands is irremediably lost when the user is de-registered. Although such a situation might sound like an exceptionally sloppy one, experience shows that this sort of practice is extensive among researchers in HPC facilities.

According to the statistics at the HLRS, as of December 1, 2017, there are 57 de-registered user accounts in the tape archive system (of a total of 262 accounts) leaving behind about 619 TB of potential dark data. In terms of percentage, 22% of the accounts (about 3.16% of the whole data volume stored at the HLRS on tape) could potentially qualify as dark data.

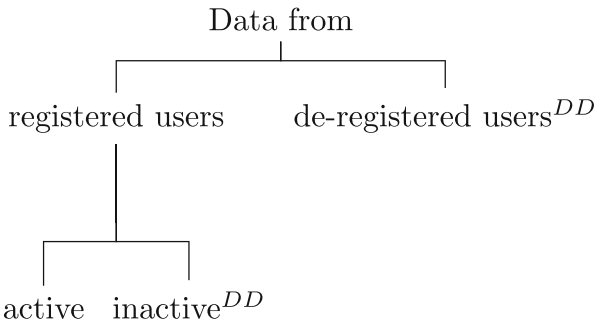

Fig. 1 Universe of users in a typical HPC facility. Users that might be the source for dark data are marked with an upper case ^{DD}

As for the *registered users*, we have to further distinguish between *active* and *inactive* users. Now, there is no consensus among HPC facilities nor recommendations by funding institutions for establishing when a registered user should be changed from an active user to an inactive user. Standard practice is that each facility determines this based on their own interest, fluency of researchers, and availability of resources. Identifying when a user becomes inactive, however, is a crucial matter for issues like dark data, since sensitive research could be lost. This situation poses the problem of determining which users, among the registered one, could potentially create dark data. To put the same problem differently, the problem is now to determine when an active user becomes an inactive one.

There are at least two standard ways to identify inactive users. One way consists in determining the time of last access of an active user to her data, that is, by keeping track of the lapse in time between the last log in and a new one. This is the preferred method of system administrators because the storage system provides such information rather easily. Now, a definition of an (in)active user attached to a temporal measurement raises some concerns. Consider the case of a user regularly accessing her files with the sole purpose of listing them. Although the user is not strictly speaking using her data, she cannot be considered inactive under the previous considerations, and thus her storage space must be maintained. Such a case is a secure avenue for dark data.

Alternatively, we could define an (in)active user by means of the "data transferred," irrespective of a lapse of time between access. By "data transfer," we refer to the size of uploads, downloads, and modifications of a file. Thus understood, a user must transfer up to a certain amount of bytes or files in order to be considered active or inactive. The main concern with this approach is that in most HPC facilities, system administrators do not usually keep track of difference in a user's data between one access and the next, the only way to measure the amount of data transferred by a given user. Another concern of this alternative is that it prompts the question, what is the lower threshold for the data transfer? Consider a user that regularly accesses, writes, and transfers her data in small amounts—below the established threshold—although very much active, she would count as inactive.

Determining when a user is no longer active depends on several factors, including the time that each HPC facility decides to give to each user. In this respect, we issue the recommendation of a five-year time period since the last access should be enough for a user to be considered inactive, and thus her data to be acted upon. This estimation is based on the fact that the lifetime for HPC systems is determined to be five years on average (Wienke et al. 2015). Computing, storage, and the users' file systems change, and unless specific projects for migrating old accounts is in place, many users and their corresponding data will be turned dark. For these reasons, it seems to us that a five-year waiting period is sufficient for any user to reclaim their data—or to decide upon them before they turn dark. Let it be noted that the DFG regulations establish a general period of ten years of storing data, allowing in this way enough time for data to be checkable and reusable by the scientific community. However, it also states that where "justified, the institution can stipulate shorter retention periods for primary data which cannot be stored on permanent and secure carriers" (DFG 2013, 75). Allowing HPC facilities to stipulate a shorter retention

Table 1 Summary of the de-registered/inactive accounts and their ratio to the overall number of accounts at HLRS

Origin	Dark accounts	Ratio
De-registered users	57/262	22%
Inactive users	72/262	27.5%
Dark account summary	129/262	49.2%

period is particularly important since they face the problem of data volume continuously growing and the allocation of important resources as a particularly pressing matter (Schembera and Bönisch 2017). Admittedly, decision criteria on which data will be stored are still an open issue in the HPC community.[5]

To our minds, then, an "active user" is one whose files are stored in the file system for a specific amount of time (v.gr. five years). It is assumed, furthermore, that those users are informed regarding the policy of data management and that they have an explicit interest in their own data not being marked as dark. Furthermore, an "active user" is one who is also transferring data within the stipulated time frame. Based on an average size of the files transferred at the HLRS, we suggest a minimum threshold of 100 MB over the five-year span. In this way, it is possible to be more justified in believing that such a user is effectively active and interested in her files. Let us finally note that the definition of active user ranges from an individual researcher producing the data to research institutions actively accessing the data.

A possible objection to our definition stems from a somehow disengaged user that only accesses the system sporadically. That is, a user that accesses her account within a period of five years but neither transfers data (e.g., there is no size modification of the files) nor gathers any sort of information about her data (e.g., list the directory, or copy a file). In such a case, the user is in fact inactive, although it would count as active. Although we believe that such a user does not really exist, a precautionary principle must be issued establishing that it is better to keep data active for such a user than performing any actions upon the data. Despite the fact that, by doing this, we might be increasing the amount of dark data, we also understand that such an outcome is preferable to any action that compromises the user and her data.

According to the records at the HLRS, as of December 1, 2017, there are 72 accounts of inactive users within the five-year time window (a ratio of 27.5%), with a total of 0.049PB of data.

A summary of the users' accounts in the archive system at the HLRS can be found in Table 1. There, it is shown that in total 49.2% of the accounts are either de-registered users or inactive users,[6] leaving behind a total of 669TB of potential dark data. Related to the 19,6PB of total stored data, this corresponds to 3.41% (see Table 2).

[5] At the HLRS, there are no registries of data older than 9 years old. This is merely because of the age of the long-term storage system, and not because data is being deleted. Any deletion is carried out by the user, and in no way is the HLRS involved in deleting users' data as of now.

[6] Several reasons explain the large number of inactive accounts. Chiefly is the fact that the HLRS has a group-based user management, creating an account for each member of a group despite such a group being composed of only one person.

Table 2 Summary of the total dark data and their ratio to the overall data volume at HLRS

Origin	Dark data	Ratio
De-registered users	0.62PB/19.6PB	3.16%
Inactive users	0.049PB/19.6PB	0.25%
Dark data summary	0.669PB/19.6PB	3.41%

Why is keeping track of de-registered and inactive users so important? As discussed earlier, such users are potential sources of dark data, and dark data has specific costs of production and maintenance. Consider first the effective monetary cost of the storage architecture. Data centers typically keep at least two copies of the data stored, resulting in a double cost of tape cartridges, libraries, and tape drives.[7] Another source of costs is the so-called total cost of ownership (TCO), which includes—besides the media cost—the energy consumption, server cost, tape drive hardware, and personnel for maintenance of the system and education. For example, the HLRS uses two copies at two physically distinct locations with different tape technology, each destined to increase the data security in case of a malfunction of the system, a fire in the building, or some other misfortune. The first copy is stored on IBM TS-1140 and TS-1155 enterprise media, being the most expensive storage media, whereas the second copy is stored on LTO6 and LTO7 media.

Finally, perhaps the most significant of all costs is of generating the data itself. This includes the concrete "dollars and dimes" perspective, where educating the researcher for producing specific data is factored in to the costs of the data themselves, along with the reusability of the data. Institutions, research facilities, and governments require a cost-effective relation between generating and maintaining data, and the accessibility to such data. In the following section, we elaborate on a new position, the scientific data officer (SDO) that aims at reducing the amount of dark data in an HPC facility by means of control and proper labeling, reducing in this way costs and fostering overall scientific productivity.

4 The Scientific Data Officer

Any attempt to diminish or even try to eliminate dark data from HPC facilities cannot be tackled exclusively with technical efforts, but a profound change in the administrative structure of how institutions manage and control their users and their data is also needed. A 2013 article published in *Nature* calls attention to "people power" and the emergence of a "new breed of scientist" (Mattmann 2013) capable of solving big data challenges by having the skills of a scientist as well as those of a computing engineer. The article calls for a new form of scientific research, reliant on computer-based technology but which conserves standard methodologies. An important aspect of this new breed of scientist is that, unless the right resources are in place, they will

[7] A good cost estimation can be found in Quantum (2018), where an overview is given of the media costs as well as the TCO of different tape technologies for 5PB of data.

be involved in carrying out many of the managerial activities that come with the new technologies, and which will distract them from their own research. Professionally, economically, and socially unmotivated to correctly label data, adapt old file formats and adopt new file formats, and overall keep data in good health, this new breed of scientist will be a major source of production of dark data, very much like their predecessors, today's researchers (Cox and Pinfield 2013).

Issues about dark data are especially pressing when researchers move away from the HPC facility. In such cases, and unless the researcher leaves successors to her work, data becomes orphaned, with no responsible individual (or team).[8] That is to say, data no longer has a responsible researcher, and it is to be expected that neither the system administrator nor any other researcher will be willing to take responsibility for such data. Orphan data becomes a "hot potato" whose most likely future is to be forgotten on storage servers. Whereas this might be the safest course of action for system administrators and researchers, it creates a significant burden on the institution in terms of costs and use of resources, as well as promoting the misplacement of scientific research.[9] Orphan data, therefore, produces more dark data. The new breed of scientists, after all, might not be free from the old managerial afflictions that affected their predecessors. For these reasons, HPC facilities have a series of decisions ahead of them: how long to keep data on storage servers, when to delete sensitive data, and who is responsible for the stewardship of the data.

We believe that these issues can be dealt with within the scope of the *scientific data officer* (SDO), a new position within HPC facilities acting as a moderator among researchers, administration, and management. In the following, we discuss what comprises the SDO and her responsibilities. We also present some practices emerging from this new position that could lead to ethical concerns. By the end of this section, we compare the SDO with other positions such as system administrator and chief data officer, and argue that the SDO has a different role and set of responsibilities within HPC facilities.

Before proceeding, let us mention that we characterize the SDO as part of the personnel within the HPC facility, and as such knowledgeable of the projects, past and present, carried out within the institution, as well as the ethical and legal framework adopted for those projects. In this respect, the SDO does not necessarily need to be a scientist or an engineer, but she needs to have training in data management, operating systems, and system administration, as well as being knowledgeable of the relevant rules and regulations affecting research, as well as rules of professional conduct. On a more general level, the SDO is responsible for the storage, curation, and general good health of data in the HPC facility (e.g., following the FAIR principles of scientific

[8] Although we will neither discuss the existence of *orphan data* in HPC facilities nor the implications for data management, it is important to be aware of this problem. Of special interest is how orphan data is incompatible with responsible research. In fact, as open science becomes a top priority in the forthcoming scientific endeavor (Cox and Pinfield 2013), orphan data needs to be adopted—or responsibly eliminated—from HPC facilities, and dark data needs to be made visible for open reuse. To support all this, we propose the introduction of the scientific data officer.

[9] By "disposable," here we mean research that produces data that, despite having epistemic value in themselves, are deleted without further assessment.

data management of accessibility, interoperability, reusability, keeping data findable, and redundancy of data (Wilkinson et al. 2016)). She is also responsible for ensuring that researchers follow standard guidelines and protocols for every project at the institution according to the FAIR principles, as well as observing and enforcing the ethical values adopted by the institution. Let us discuss these responsibilities in more detail now.

4.1 Technical Responsibilities

We identify five technical responsibilities for the SDO. Those are as follows.

First, labeling—or supporting the labeling of—data with standardized metadata according to the FAIR data management principles. This labeling provides several advantages, such as allowing fast location of data, exchange of data among researchers, and maintaining overall control of computing resources and storage. Labeling data with standardized metadata is the cornerstone for reducing dark data in HPC facilities. Let us note that it also facilitates research programs—such as those promoted by the open science movement—by standardizing data across projects, institutions, and computational systems. In this respect, the SDO must be able to quickly incorporate domain-specific information into existing metadata schemata, as well as to develop a metadata schema from scratch if necessary. One possible general metadata model might come from the Open Archival Information System standard, and has to include descriptive—this part incorporating the domain-specific metadata—administrative/preservation, and format metadata (OAIS 2012). We will have more to say about this later.

If the new breed of scientists are to perform better than their predecessors, the very first step is to motivate their work by reducing the largely unrecognized job of labeling their own data. For this task, the SDO can contribute a great deal.[10] In this respect, the SDO would be responsible for the creation of metadata in agreement with the owner of the data, ensuring that data is labeled according to standardization and following FAIR principles, and accountable for securing the data and metadata on the servers for easy localization and retrieval.[11]

Given full access to data and metadata, there is the possibility that the SDO could assist researchers in several ways. For instance, the role of SDO is key for projects intended to reuse data. Indeed, the SDO is the sole position that brings together all

[10]Transferring the job of labeling data to the SDO is a problem in its own right that we cannot address here. However, we understand that there are at least two issues to be discussed. First is the question of how will the communication between researchers and the SDO be implemented. The SDO by herself cannot label data without a minimum of information exchange with the researchers involved. The second issue is the role of the researchers in the new context of labeling data. While our provisions are meant to motivate scientific research in HPC facilities, we are aware that it could also foster the researcher's lack of concern about their data, and her subsequent neglect of the responsibilities that comes with it. In this respect, technical, ethical, and legal responsibilities cannot be shifted from the researcher over the SDO, but rather must be shared by all parties—including, of course, the institution itself.

[11]A topic arises here which is not initially covered by the FAIR guidelines, namely, that data stewardship assumes responsibility for the data (Wilkinson et al. 2016). To us, the SDO must be responsible for the visibility of the data, which includes the periodic checking of the data inventory and the stewardship over data of individuals that have left the HPC facility (Hick 2010).

projects currently running, and which were active at the HPC facility at a given time. Furthermore, one could also anticipate that, at an inter-institutional level, researchers would share information and data through the SDOs of each institution. In such a case, it would no longer be the reputation of the scientist—or of the institution—that allows research data to be shared between institutions, but rather a network of SDOs that vouches for any and all researchers of a given HPC facility.[12] Another important way in which the SDO assists researchers is for cases of reproduction and replication of results of computer simulations. In such cases, the SDO assists in retrieving previous data of computer simulations of potential interest for the researchers; another way to assist is through intra- and inter-institutional consultation of whether the intended computer simulation is viable (e.g., regarding ethical matters, a history of (un)successful simulations. See Durán and Formanek (2018)). Admittedly, replication and reproduction are a time-honored problem in science and engineering of which we have no intention to discuss in length. However, it is important to notice that these problems acquire a sense of urgency in active projects at any HPC facilities, where methodological, technological, and economical constraints restrict the possibilities of obtaining comparable data (e.g., not every institution can reproduce the results of a project in an HPC facility. Barba (2016) and Barba and Thiruvathukal (2017)).

Second, if the institution runs its own data management systems, the technical administration and support for these systems may also be the domain of the SDO. In such a scenario, the specifications for the position is that the SDO must be acquainted with recent trends in research data management, HPC-related tools, and modern HPC infrastructures and has working knowledge of in-house data storage facilities. Admittedly, many of these tasks overlap with other positions within the HPC facility, such as the system administrator. However, and as we explain in Section 4.3, the SDO plays a different role within the institution, despite some overlapping obligations and responsibilities. Of course, a further discussion is needed to determine whether there is an explicit hierarchy, and how duties and responsibilities that overlap will be resolved.

Third, within an organizational unit, the SDO has to act as a multiplier and transfer knowledge to all members of the institution, whether they are involved in a given project or not. This means that communication among researchers should also go through the SDO, who could in turn contribute with her own understanding of the managerial problems (e.g., of whether certain data could be used for the validation of computer simulations). She must also provide training and support to researchers, such as in the production of metadata and their need for its future reusability, reproducibility, and replicability. It is important to note that such training does not aim to give managerial tasks back to the researchers, but rather to facilitate the labeling process that will ultimately make the right data available for the intended research.

Fourth, the SDO must be responsible for the visibility of data as a measure for reducing—and eventually eliminating—the presence of dark data. This task also

[12] Let it be noted that this context also suggests that the SDO is not a position exclusively for HPC facilities, but rather for any institution involved in heavy computer-based research.

includes periodic checking of the data inventory for potential dark data by de-registered or inactive users. When an inactive user is detected, the SDO should contact the respective user in order to clarify if the data can be deleted or if the user still has interest in the data. The SDO has a specific weight in deciding on the time a user is considered active in the system. This decision, as discussed earlier, is not trivial and affects all parties: in-house researchers, the institution, research foundations, and the flow of research itself. It is advisable, therefore, that researchers, the institution, and the SDO—in the role of the expert—all agree regarding such decisions.

Fifth, the SDO also has to take over the stewardship of dark data. When data becomes dark either by de-registered or inactive users, the SDO has to fill the lack of a responsible researcher and take the data under her purview. Given that data stewardship is not initially covered by the FAIR principles, we strongly recommend that the SDO becomes directly responsible for the data from those users that have left the facility and which have not been properly labeled. Shifting responsibilities for the data, we believe, is legally and ethically possible, and overall it is the best action against dark data.

We believe that the SDO has the capacity to implement the kind of decision criteria needed for the stewardship of data broadly conceived. Such criteria must work as the framework within which the SDO can act and make decisions on sensitive matters such as determining when data must be deleted from the system to free resources, which data may be deleted, and how to act in case of conflict (e.g., between researchers and among competing projects). Although the SDO takes the role and responsibility of implementing such decision criteria, it is very important to balance the control and authority over data with other actors and stakeholders. In this respect, the decision criteria used for data must have been previously agreed upon with the researchers, discussed with the head of each department—and any internal commission—and included in the terms of agreement for the use of the servers. Special commissions should also be formed with the aim of standardizing the decision criteria across facilities and which constitutes the common base for all SDOs.

Finally, allow us to suggest a decision criteria for data storage and maintenance in HPC facilities. As the amount of data grows, it is obvious that it is not possible to store all the data despite the efforts of institutions to keep up with new technology. Because of this, data has to be carefully selected according to the impact/importance in relation to volume. This relation is depicted in Fig. 2, where "publication with plots" (tier 4) is highlighted as the most important scientific object in research. Typically, these are preserved in libraries and therefore not the focus of our work. Instead, the tier of interest is the input data, the simulation code, and the results of the evaluation (tier 3). Selected simulation results (tier 2) are also important to us; however, they depend on the amount of data. Of less importance are the unevaluated simulation results (tier 1), test runs and intermediate results, which usually can be deleted without much loss. It is expected, however, that each institution finds an equilibrium between the importance of their data and their volume according to their own interests, resources, and aims. In this respect, the data pyramid in Fig. 2 should only be taken as a suggestion.

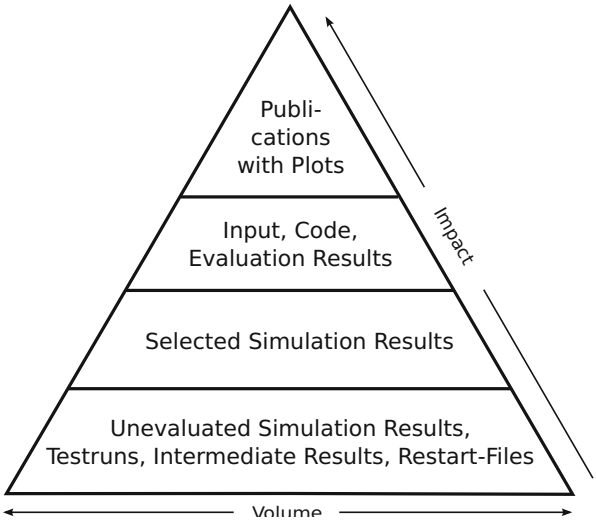

Fig. 2 Data pyramid for simulation data showing the relation between impact of the data and volume. According to this relation, data has to be selected (see Reilly et al. (2011) and Iglezakis and Schembera (2018))

In order to comply with the above technical considerations, it is important that the SDO has a good understanding of the data and metadata, as well as be acquainted with past and current projects at the HPC facility. It is also advisable that the SDO maintains a network of peers across HPC facilities and other institutions using or producing relevant data, for in this way, scientific collaboration can be significantly fostered. Given the place that the SDO has in an HPC facility, it should be expected that the SDO holds key positions on consultation boards, technical committees, and ethical panels.

4.2 Ethical Considerations

The SDO's responsibilities go beyond managerial and technical duties, as she must also observe and enforce good scientific and engineering practice. In this context, two considerations emerge. On the one hand, the SDO must be vigilant of the researcher's professional behavior and enforce, when necessary, the appropriate reprimand or legal action. In this respect, many standard ethical concerns about professional behavior and research practice are at play: misappropriation of research results, interference in the production of research results, obstruction of scientific and engineering research, noncompliance with codes, among others. On the other hand, there are moral considerations regarding the control and authority of the SDO position itself. To this end, the SDO must not only be vigilant of her own behavior but also considerations over limiting the SDO must be discussed. In this context, there are at least two chief issues of significant moral concern that we will be addressing, namely, overseeing and controlling the user and her data and placing data in the servers. Let us discuss each one in turn.

Earlier, we maintained that each HPC facility is responsible for determining when a user is active or inactive, examining this with the case at the HLRS. Making such decisions over the users not only affects them but also their data, potentially rendering them dark. An effective solution is to adopt an SDO as chief account and data manager, instead of adopting automatic scripting that activates/inactivates users and their data. The reason for choosing the SDO over automation is that special circumstances require a deliberate approach. For instance, we mentioned that a user is active while she is within the five-year span. Consider now a user that has no activity on her account for the large part of four years, but right before the turn of the fifth year she accesses the system by simply copying her files to a different folder within the institution. Suppose further that this happens on a regular basis for long periods of time. Evidently, this user is interested in keeping her data stored and her account still accessible—at the cost of multiplying data. Within an automatic scripting system, the user's inactive time will be reset and her data maintained within the servers. With the SDO, however, the situation changes because she could evaluate with more sensitivity the factors that determine when a user becomes inactive, and justify whether it is appropriate to maintain such a user as active or not, and therefore maintain or delete her files from the servers. Among these factors, we count the SDO responsible for considering the time and frequency of access to the user account and data (i.e., the number of accesses within the five-year span and the overall accesses since the first time the user was active), the type of access to the data (e.g., listing the directory is not the same from the point of view of the HPC facility as copying data), the reliability of such data (e.g., by reports from other users that the data is not suitable for certain purposes), the compatibility of the data (i.e., their format) with respect to other projects related to that HPC facility and other institutions, among others.

The SDO also works as a more efficient gatekeeper of the data than an automatic scripting against the misbehavior of users. The SDO is able to determine the intentions of the user (e.g., a researcher no longer affiliated to the HPC facility copies data to another institution evading in this way protocols of data sharing) and evaluate the costs for the institution (e.g., in terms of resources, research efforts, and the value of data[13]) by accounting for the amount of data transferred, to which location, and within which frequency. Equally important is the type of operation performed with the user's data. Copying data could constitute an infringement or professional misbehavior if the user is no longer associated with the HPC facility, but deleting the data from the servers without the proper authorization could amount to civil and even criminal charges. It is essential then that HPC facilities adopt the SDO as their most effective control over the users and their data as well as a protective measure of the public good that is scientific and engineering research. Thus understood, the role and responsibility of the SDO cannot simply be replaced by an automatic system, but it

[13] Admittedly, it is difficult to put a value on data. Some data are more valuable than others considering the effective cost of production (e.g., computational power, human resources involved). Some other data have more value given the target system they represent (e.g., data about floods is more valuable to certain regions than data about climate change, despite the former being obtained at lower costs than the latter). Yet, some other data are more valuable considering the efforts that a country or an institution puts into them (e.g., low technologically driven societies having accurate data about the health of their population).

needs to be carefully assessed by an authority. In this sense, the visibility of the SDO in an HPC facility is also important for the good functioning of the institution, for now the researcher's interests and concerns converge into one visible position.

Unfortunately, the SDO also has a great deal of control and power over the users and their data, along with an unquestionable overseeing authority within the institution. Following on the previous situation when the user leaves the HPC facility, the SDO needs to decide whether such a user should be marked as inactive and, if that is the case, what to do with her data. On the one hand, the SDO must prevent the data from becoming dark, and therefore they must be appropriately stored; on the other hand, marking a user as inactive could lead to a series of demands from the user, as she is vulnerable to the decisions of the SDO—after all, the user was accessing the system regularly within the five-year time window. HPC facilities need to anticipate these cases and create a legal framework for where to situate the SDO's activities. Such framework, it must be noticed, is meant to protect not only the SDO but also the institutions, the users, and their data.

Equally concerning is the overarching control of the SDO over the users and their data that are still affiliated and working at the HPC facility. Consider the case of an SDO that favors competing projects by allowing one group of researchers to retrieve data before any other group, or showing favoritism by labeling the first group's data faster than other groups. Such abuses of power must be limited. In the same vein, the SDO has the ability to object to the data produced by one group of researchers based on lack of interest to the HPC facility and its general research aims. The individual and subjective character of these decisions could lead to all sorts of conflicts. Furthermore, since HPC facilities typically have different storage media, the SDO could grant storage permission to one group on more secure media—ensuring in this way that research will be preserved in the long-term—whereas the second group stores their data on regular, less reliable media. In such a case, the former group will keep their research protected while the latter group exposes their data to degradation over time. Such a situation creates visible favoritism and disadvantages between research groups and it must be prevented. One last concern is that the SDO has the capability of deleting users' data from storage without any further control by any other branch of the institutional administration. The disappearance of data from the servers is an infraction of basic principles of scientific and engineering practice, and justifies a *prima facie* assumption of dishonesty and gross negligence which lays entirely in the hands of the SDO.

A first step towards avoiding these cases is that all workflow directed to the SDO must be in accordance with the GDPR regulations (GDPR 2016).[14] Unfortunately, these are only initial attempts to limit the control and authority that the SDO has over users and their data. A further preventive measure could include the distribution of authority between the SDO and special committees—either regulating the SDO's activity or actively involved in sensitive issues such as deleting data and granting equal access to all researchers.

[14]This is for HPC facilities such as the HLRS that must comply with the European Union data regulations. Other countries should adopt their own statutes.

Fabrication and falsification of data, two chief concerns in scientific and engineering research, also affect the competence of the SDO. As it is known in the literature, fabrication of data consists in researchers deliberately creating data with the purpose of propounding with greater conviction a particular hypothesis. Falsification of data, on the other hand, occurs when data is altered or omitted in order to fit in with a hypothesis. Both instances constitute serious cases of professional misconduct and should be seriously considered and penalized. We submit that the SDO is in a position of power and control that could also lead her to forms of fabrication and falsification of data even when she is not the producer of such data. To be more specific, fabrication and falsification of data come in the form of *misplacing data* in the servers, by means of wrong, incomplete, or misleading metadata. The first two forms of misplacing data (v.gr., wrong and incomplete metadata) result in data moved away from the right database. In this way, the SDO can deliberately create a database where data that do not fit the desired hypothesis are absent (i.e., fabricates data). Misleading metadata, on the other hand, consists in complete but false information about the data. By means of deliberately modifying the metadata, the SDO can cluster data in specific groups, altering the results of research based on statistical data (i.e., falsification of data).[15] In truth, misplacing data could be the work of intentionality and premeditation, just as much as it could be the result of a negligent SDO. In either case, we submit that this sort of behavior constitutes an infringement of any codes of professional conduct and must be penalized.

In truth, misplacing data is perhaps the easiest way to detect when the SDO misbehaves. Other related ethical issues that have a kinship with misplacing data are forms of *data selection and manipulation*. After results have been obtained from computer simulations, the SDO has the control to mark data as available, disposable—due to the fact that not all data needs to be stored after an experiment or test—and even withhold it until further evaluation is carried out (e.g., to evaluate whether data fulfills GDPR standards). The last situation allows the SDO to withhold data from researchers under the pretext of further evaluation and assessment, affecting in this way individual research as well as favoring competing research groups and institutions. Recall from our previous discussion how the SDO could favor one group of researchers over another by attending one group's needs over the others. The situation here is analogous with the exception that such favoritism is carried out via the selection of data rather than the selection of the researcher. Another simple mechanism that allows the manipulation of data consists in influencing the researcher's decisions by the way data is presented. In such a case, the researcher must decide on whether data is apt for reuse—reproduction, replication—or their research requires gathering further data by running new computer simulations. These considerations make evident the oversight that the SDO has over the researchers, the data, and research programs at the HPC facility. In these particular cases, the SDO is intentionally degrading the intrinsic value of data, potentially increasing the amount of dark data, and mislead scientific and engineering research.

[15] Admittedly, misleading metadata could also be used for fabrication of data.

In view of this, the control and authority of the SDO must be regulated and limited. One way of doing so is to involve the research community and create instances of auditing the SDO's decisions and actions. Thus, researchers should be able to file complaints when they feel mistreated, persecuted, or victimized in any way. Another way to limit the control and authority of the SDO is to involve the researchers in-house along with some administrative branches at the HPC facility (e.g., those responsible for funding, the director of the institute, heads of departments) in decisions regarding sensitive data (e.g., data with a high cost of production, results that could be controversial). In an ideal democracy, no one individual or position in the public sphere should hold absolute power over researchers, data, and the future of science and engineering research.

More generally, the SDO must comply with the same ethical principles and codes of professional conduct as her fellow researchers and public servants. While there are several codes of conduct available, especially in engineering, there is one that has special value for an HPC facility such as the HLRS where computer simulations are the main source of data. Tuncer Ören and collaborators have offered the most elaborate and thorough account of a code of ethics for computer simulationists (Ören et al. 1998, 2002). According to Ören et al., the fundamental problem is that the use of computer simulations might lead to serious implications for human beings as well as humanity as a whole. His chief concern stems from the fact that computer simulations are used to support crucial policies and decisions that might alter our current life and constrain our future. In nuclear fuel waste management systems, for instance, computer simulations are used to study the long-term behavior, environmental and social impact, and the means for containing nuclear fuel waste. Similarly, computer simulations of leaking nuclear waste into the Earth and underwater rivers are crucial for supporting decisions on whether such nuclear waste should be buried instead of building a special facility for storage (Ören 2002). We believe that this code of conduct is suitable for the SDO, chiefly because her work requires understanding what constitutes professional behavior within the practice of computer simulations, and its connection to the use of data for policy making, public communication of results, and scientific collaboration.[16] In this respect, good scientific practice based on large amounts of computer-simulated data heavily depends on the honesty, dedication, and proper conduct of the SDO, just as much as the researchers that produce such data.

4.3 What the SDO Is Not

Equally important to understanding the role and responsibilities of the SDO is to determine what this position does not comprise, especially in comparison with already existing positions within an HPC facility. In this respect, the SDO is a novel appointment and we claim that it has no precedent in data management positions. To show this, we compare the SDO with known positions, namely, system administrator, chief data officer, and security officer.

[16] For an overall study on the ethics of computer simulations, see Durán (2018, Ch. 7).

The SDO differs from the work of a system administrator mainly in its objective and approach. The chief aim of system administrators is the technical management of the entire storage system in a facility. In this sense, the system administrator is responsible for running, installing, patching, upgrading, and the general maintenance of systems. The chief aim of the SDO is rather non-technical, since it is about managing data in such a way that it is visible for other researchers, and maintaining data for reuse, reproduction and replication purposes, among other responsibilities.

An interesting fact about the position of a system administrator is that they are "invisible" for the users, since their job is the maintenance of the system. The SDO, instead, is constantly related and involved with the researchers of the HPC facility, and as such her work requires special social skills: communicating changes in laws and regulations; planning and budgeting new equipment acquisitions; acting as an intermediate between the user's needs, the available resources (inter- and intra-institution), and fulfilling the general objectives of the HPC facility; among other social skills. There is, however, some technical work also associated with the SDO. As discussed, the SDO is in charge of the maintenance of the data, the proper labeling with metadata, and other technical aspects. In this respect, the SDO will also interact with the system administrator in constructing a reliable infrastructure of information services. While there is much literature on the role of a system administrator (e.g., on UNIX system administration (Nemeth and Whaley 2015; Bergstra and Burgess 2011)), there is little discussion on the limits of this position and what it means for the functioning of an HPC facility (Mann et al. 2003).

The SDO also differs from the chief data officer (CDO), the latter understood as responsible for determining what kinds of information a company chooses to capture, retain, and exploit and for what purposes (Lee et al. 2014).[17] One meritorious difference is that the CDO approaches problems from a business perspective, reflecting specific issues in business contexts (e.g., how to implement the CDO as a new executive among chief information officer (Lawry et al. 2007; Peppard 2010), chief technology officer (Smith 2003), and others), rather than focusing on issues stemming from scientific and engineering research. Furthermore, our approach proposes a different organizational hierarchy as the SDO is involved in management, communication, and administration. In contrast to the CDO, there is neither a top-down nor a bottom-up hierarchy within its role but rather a supportive, responsible authority.

Admittedly, of all existing positions related to data administration in HPC facilities, the SDO is perhaps closest to the security officer (SO), who is responsible for all areas involving cybersecurity, including the technical, management, and legal areas (Whitten 2008). On the technical side, the SO deals with detecting unauthorized intrusions, applies the latest security patches, and keeps updated firewalls, anti-virus, software of various kinds, etc. Other tasks for the SO are disaster recovery planning, maintaining vendor relations, and coordinating legal actions to be taken with respect

[17] In a company or governmental institution, there are positions similar to the CDO, such as the chief digital officer (CFO) and the chief digital information officer (CDIO). The work of CFO and CDIO is to drive the company or institution into the digitalization of their data, modernize their infrastructure, and oversee operations over mobile applications, social media, and the like. None of these roles are ascribed to the SDO.

to security incidents. Since cybersecurity is at the intersection of technical, legal, and organizational topics, the position of the SO has to cover a lot more than conducting technical work, generally invisible behind the scenes. She has to actively undertake actions and become visible in various forms: as a contact person, as the proxy with SDOs in other institutions, and is responsible for the data and their storage. Visibility and multidisciplinary are also key characteristics of the SDO, though only with respect to data and not systems security.

5 Final Remarks

Our efforts have been focused on two core issues. First, to conceptualize dark data in HPC facilities as well as to show that it is the product of common practice among researchers. To this end, we discussed some current statistics at the HLRS, a major HPC facility in Europe. The fact that dark data is invisible should not suggest that it is non-existent. The second aim was to present and characterized the scientific data officer as the new position in HPC facilities that can deal with issues emerging from data in general, and dark data in particular. In order to show this, we discussed the SDO's technical and ethical responsibility, as well as distanced the SDO from standard positions such as system administrators, CDO, and the SO.

A key element in understanding the sources of dark data is that it enables the design and planning of containment measures in HPC facilities, such as the SDO and the personnel and infrastructure that comes with this position. Furthermore, data is a scientific and engineering commodity of high public value. For this reason, it is pressing that data can be reused in research as well as be available to the whole scientific community. The SDO position is explicitly designed to help in this endeavor, promoting the responsible use of data, increasing workflow, helping to cut down unnecessary costs (e.g., the costs of producing and maintaining already available data), and of course bolstering the implementation of standards such as FAIR and those ascribed by the open science movement (Suber 2012).

Acknowledgments We thank the High-Performance Computing Center Stuttgart for their support during this research. Juan M. Durán also thanks the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) for their support. Finally, both authors thank Johannes Lenhard for valuable comments on an earlier version.

References

Austin, B. (2014). Dark data: what is it and why should I care? https://www.r1soft.com/blog/dark-data-what-is-it-and-why-should-i-care, last visited 2019-02-19.

Babcock, C. (2015). IBM Cognitive colloquium spotlights uncovering dark data. Information Week. https://www.informationweek.com/cloud/software-as-a-service/ibm-cognitive-colloquium-spotlights-uncovering-dark-data/d/d-id/1322647, last visited 2019-02-19.

Barba, L.A. (2016). The hard road to reproducibility. *Science*, *354*(6308), 142–142.
Barba, L.A., & Thiruvathukal, G.K. (2017). Reproducible research for computing in science & engineering. *Computing in Science & Engineering*, *19*(6), 85–87.
Barberousse, A., & Marion, V. (2013). Computer simulations and empirical data. In Durán, J.M., & Arnold, E. (Eds.) (pp. 29–45). Computer simulations and the changing face of scientific experimentation: Cambridge Scholars Publishing.
Bergstra, J.A., & Burgess, M. (2011). *Handbook of network and system administration*. New York: Elsevier.
Brantley, B. (2015). The API briefing: the challenge of government's dark data. https://digital.gov/2015/06/17/the-api-briefing-the-challenge-of-governments-dark-data/, last visited 2019-02-19.
Choudhury, S., Fishman, J.R., McGowan, M.L., Juengst, E.T. (2014). Big data, open science and the brain: lessons learned from genomics. *Frontiers in Human Neuroscience*, *8*(239), 239. https://www.frontiersin.org/articles/10.33/fnhum.2014.00239/full, last visited 2017-12-03.
Cox, A.M., & Pinfield, S. (2013). Research data management and libraries: current activities and future priorities. *Journal of Librarianship and Information Science*, *46*(4), 1–18.
Darch, P.T., & Sands, A.E. (2015). Beyond big or little science: understanding data lifecycles in astronomy and the deep subseafloor biosphere. iConference 2015 Proceedings. https://www.ideals.illinois.edu/bitstream/handle/2142/73655/185_ready.pdf, last visited 2019-02-20.
Dennies, P. (2015). Factories of the future: the value of dark data. Forbes BrandVoice. https://www.forbes.com/sites/teradata/2015/02/19/factories-of-the-future-the-value-of-dark-data/, last visited 2019-02-20.
DFG (2013). Proposals for safeguarding good scientific practice recommendations of the Commission on Professional Self-Regulation in Science. Tech. rep., Deutsche Forschungsgemeinschaft, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf, last visited 2019-02-20.
Durán, J.M. (2018). *Computer simulations in science and engineering. Concepts - practices - perspectives*. Berlin: Springer. https://doi.org/10.1007/978-3-319-90882-3, ISBN 978-3-319-90880-9.
Durán, J. M., & Formanek, N. (2018). Grounds for trust: essential epistemic opacity and computational reliabilism. unpublished.
Edwards, K., & Gaber, M.M. (2014). *Astronomy and big data. A data clustering approach to identifying uncertain galaxy morphology*. Berlin: Springer. https://doi.org/10.1007/978-3-319-06599-1.
GDPR (2016). Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Parliament and the Council of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj, last visited 2019-02-20.
Glass, R., & Callahan, S. (2014). *The Big Data-driven business: how to use big data to win customers, beat competitors, and boost profits*. Berlin: Wiley.
Goetz, T. (2007). Freeing the dark data of failed scientific experiment. *Wired Magazine*, *15*(10), 7–12. http://www.wired.com/science/discoveries/magazine/15-10/st_essay.
Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299.
Heidorn, P.B., Stahlman, G.R., Steffen, J. (2018). The astrolabe project: identifying and curating astronomical 'dark data' through development of cyberinfrastructure resources. *Astrophysical Journal Supplement Series*, *236*(1), 3. https://doi.org/10.1051/epjconf/201818603003.
Hick, J. (2010). HPSS in the Extreme Scale Era: report to DOE Office of Science on HPSS in 2018-2022. Tech. rep., Lawrence Berkeley National Laboratory, https://escholarship.org/uc/item/4wn1s2d3#main.
Humphreys, P.W. (2013). What are data about? In Durán, J.M., & Arnold, E. (Eds.) *Computer Simulations and the Changing Face of Scientific Experimentation*: Cambridge Scholars Publishing.
Iglezakis, D., & Schembera, B. (2018). Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart-Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING. *o-bib Das offene Bibliotheksjournal/Herausgeber VDB*, *5*(3), 46–60.
Lawry, R., Waddell, D., Singh, M. (2007). Roles, Responsibilities and futures of chief information officers (CIOs) in the public sector. *Proceedings of European and Mediterranean Conference on Information Systems 2007*.
Lee, Y., Madnick, S.E., Wang, R.Y., Wang, F., Zhang, H. (2014). A cubic framework for the chief data officer: succeeding in a world of big data. https://dspace.mit.edu/bitstream/handle/1721.1/103027/esd-wp-2014-34.pdf?sequence=1.

Leonelli, S. (2013). Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. *Bulletin of Science Technology & Society*, *33*(1-2), 6–11.

Leonelli, S. (2014). What difference does quantity make? *On the epistemology of Big Data in Biology, Big data & society*, *1*(1), 1–11.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821.

Mann, M., Sachs, W., Aschemann, G., Krabbe, G., Austermühle, S., Jellinghaus, A. (2003). Gedanken zum Berufsbild des Systemadministrators - Diskussionsgrundlage für sage@guug. http://www.guug.de/sage/texte/berufsbild-2003-01-10.pdf, last visited 2019-02-20.

Mattmann, C.A. (2013). *Computing: a vision for data science* Vol. 493: Nature Publishing Group.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Nemeth, E., & Whaley, G.S.T.R.H.B. (2015). *UNIX and Linux system administration handbook*, 4th. Vol. 7: Addison Wesley.

OAIS (2012). *Reference model for an open archival information system (OAIS), recommended practice. CCSDS 650.0-M-2 (magenta book) issue 2 consultative committee for space data systems*: Washington.

Ören, T.I. (2002). Rationale for a code of professional ethics for simulationists. *Summer Computer Simulation Conference*, 428–433.

Ören, T.I., Elzas, M.S., Smit, I., Birta, L.G. (1998). Code of professional ethics for simulationists. In *Summer Computer Simulation Conference, Society for Computer Simulation International* (pp. 434–435).

Ören, T.I., Birta, L.G., Elzas, M.S., Fairchild, B., Smit, I., i Erols, M.A.P. (2002). Code of professional ethics for simulationist. Society for Modeling and Simulation International. http://scs.org/ethics/, last visited 2019-02-20.

Peppard, J. (2010). Unlocking the performance of the chief information officer (CIO). *California Management Review*, *52*(4), 73–99.

Quantum (2018). Quantum White Paper. LTO: The new "Enterprise Tape Drive". http://www.quantum.com/iqdoc/doc.aspx?id=15146, last visited 2019-02-19.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M. (2011). *Report on integration of data and publications*. Tech. rep., Alfred-Wegener-Institut, http://epic.awi.de/31397/1/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf, last visited 2019-02-20.

Schembera, B., & Bönisch, T. (2017). Challenges of research data management for high performance computing. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries* (pp. 140–151). Cham.: Springer.

Shahzad, M.A. (2017). *The big data challenge of transformation for the manufacturing industry*. IBM Big Data & Analytics Hub. http://www.ibmbigdatahub.com/blog/big-data-challenge-transformation-manufacturing-industry?, last visited 2019-02-19.

Smith, R.D. (2003). The chief technology officer: strategic responsibilities and relationships. *Research-Technology Management*, *46*(4), 28–36. https://doi.org/10.1080/08956308.2003.11671574.

Suber, P. (2012). *Open Access*. MIT Press: MIT Press.

Whitten, D. (2008). The chief information security officer: an analysis of the skills required for success. *Journal of Computer Information Systems*, *48*(3), 15–19.

Wienke, S., Iliev, H., an Mey, D., Müller, M.S. (2015). Modeling the productivity of HPC systems on a Computing Center Scale. In *International conference on high performance computing* (pp. 358–375). Cham.: Springer.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, https://www.nature.com/articles/sdata201618.