Managing Big Data Issues Within a Research Data Repository:

Dealing With the 21st Century Data Explosion

James W. McNally
NACDA Program on Aging
Inter-university Consortium for Political and Social Research (ICPSR)
University of Michigan
Ann Arbor, Michigan, USA
jmcnally@umich.edu

***Abstract*— Increasingly, organizations in both the public and private sector who fund the collection of research information expect and in many cases mandate the public sharing of these data as a condition of support. While representing a positive expression of the idea that information represents a public good, it has also resulted in a veritable flood of new studies, surveys and administrative records entering the public domain; the emergence of the Big Data model in the secondary analysis of research information. Much of these data are managed by data repositories that ingest, process, clean and enhance these files so they are accurate and consistent and introduce as little error as possible into the analysis stream of information. Conversely, this huge influx of Big Data resources has also resulted in higher expectations for the rapid release of data that complicates the need for a thorough review and cleaning of files before distribution. This paper reviews emerging approaches within a research data repository that seek to maintain high quality control while managing Big Data streams as they enter the system.**

Keywords-component; Big Data, repository, distribution data error, processing

I. INTRODUCTION

While all data are new when first generated or captured, it immediately becomes historic the moment it is released and the role of managing these data goes from one of creation to one of preservation and maintenance. In the past, the lag between creation and release could be measured in years as the origin data was examined, prepped and analyzed by primary investigators. Major federal data collection efforts in the United States such as the US Decennial Census or the annual National Health Interview Survey (NHIS) routinely faced lags between the collection and release of these data of three or more years up until the early part of the 21st century [1]. That time lag between data collection and data release was often critical; rapid changes in small area populations, new technology, political shifts or new medical interventions could make the data obsolete before they became generally available to the user community. Still, the user community (academic, business, and industrial) accepted this as the inevitable pace of information release and accordingly factored this lag time into their own information needs and analysis plans. This historic pace of data release combined with the relative scarcity of data resources by today's standards created a dynamic which allowed data to have monetary value. In the late 1990's and early 2000's, for example, the National Technical Information Service (NTIS) in the US routinely sold the NHIS data at a cost of over a thousand dollars per copy, largely restricting access to these resources to well-funded organizations and research universities. While attempts to monetize federal information resources remain common, research data, particularly information supported by federal research dollars are seen increasingly seen as "public goods"; something to be made available at little to no cost for anyone with an internet connection. In less than a decade, the transition from media based distribution (tape, CDROM, USB and drive storage) to one of digital or "streaming" data transfer has completely transformed concepts of data availability and largely eliminated the accepted "waiting period" that allowed the time for new data to be fully processed before it was distributed for public consumption and reanalysis. As researchers, we increasingly live within a world where data is expected to be both free and available almost immediately after its creation.

The increased level of availability has ushered most data users into the age of what is informally called "Big Data". This term is loosely applied to a staggering number of data collection streams ranging from health information, to consumer purchases, to economic trends, to twitter feeds, and its use is attributed to a wide array of organizational units ranging from business applications to medical records to the governmental collection of cell phone use. As a consequence, there is no clear or universally accepted definition of what Big Data is other than the desire, if not the ability, to manage large arrays of data in real time. While no one group or discipline can be said to own the concept of Big Data, everyone involved in the management of information has a stake in working effectively with or within the Big Data environment.

This paper looks at the issues of Big Data from the perspective of a data repository that manages an ever growing stream of electronic research data in the form of surveys, censuses,

and administrative data. The NACDA Program on Aging is part of the Inter-university Consortium for Political and Social Research, the largest social science data repository in the world. NACDA is primarily responsible for the National Archive of Computerized Data on Aging, a collection of 1,600 plus studies on aging and health. These data are acquired, processed, enhanced and then reissued to the international research community for reanalysis, journal articles, grant development and policy applications. The role, responsibility and tasks faced by a digital repository has changed dramatically in recent years and this paper reviews the ways in which current practices are being revised and redefined to meet the growing demand for rapid and low cost information sharing across the research community.

II. BACKGROUND

A. Development of Big Data as a Construct

The emergence of the idea of "Big Data" has greatly expanded in the past two decades, but the concept has existed since the 1940's when researcher first began to think about the expanding size of the printed word. Much like the digital tidal wave of today, libraries, archives and paper storage facilities of the time were seeing a huge growth in the number of printed items (book, bill and records) being acquired by their systems. While the concept of digital storage remained largely unknown, library sciences were gearing up for a huge increase in written information that would have to be catalogued, indexed and supported by a much larger human workforce [1]. Even with the introduction of early digital storage systems such as UNIVAC in the 1950's and its subsequent rapid growth, data storage remained a clumsy, uncertain and time intensive process with no reliable form of data compression emerging until the late 1960's [2]. During the 1970's and 1980's much of the discussion addressing computer storage and information management returned to the concerns seen in the 1940's over whether or not we could effectively use the huge amount of information then available. Whether or not an unlimited amount of data could be stored became less of an issue by the 1990's, and concerns centered around the kinds of software we would need to replace human review by mechanizing the analysis and extrapolation of information culled from an ever increasing flow of information [3]. This was also a time when we began to seriously consider our growing ability and possibly our obligation to capture and store "everything" in the hopes that we would eventually extract useful information with new approaches and improved software [4].

The rapid decline in both the cost of storage and of memory and the growth of the internet were the forces that really made the age of Big Data possible. In the mid-1990's, IBM announced it had become cheaper and more efficient to store information digitally than on paper [5] and the pace of change has only increased as Moore's law has asserted itself again and again. It has now become routine for individual researchers to carry terabyte drives in their briefcases, something that would have been prohibitively expensive only a few years ago. It was in this environment beginning in the early 21st century that the term Big Data and its underlying concepts were more formalized by research presented by Diebold [6] in 2000 and then Laney [7] in 2001.

Since the heady years of the early 2000's an increasing number of researchers and institutions have begun to grapple with the scientific [8], cultural and ethical [9] [10] and technical challenges [11] that Big Data represents in terms of the successful and efficient management of these growing and ongoing streams of new and historic information being ingested by computer systems worldwide. All of these concerns are strongly felt within the digital research repository sciences as the quality, ethical use and equitable distribution of data represent the foundations of our discipline. The rapid increase in the quantity of available data has serious implications for our ability to efficiently manage this flow of often vital information from the primary data collector to the secondary user. This is a particular challenge as the quality and provenance of these new information streams can vary tremendously and often require an additional investment of repository time and resources in vetting the data consistency and performing essential confidentiality reviews to insure the protection of study respondents.

B. Archival Perspectives on the Growth of Big Data

From the perspective of the archivist working in a research data repository, the transition from tape storage to the routine use of the spinning disc and cloud based services connected to the internet as a primary storage medium represents the technology shift that has driven the rapid change in the ability to move large quantities of data quickly and safely from producer to user. This transition in delivery access has had both positive and negative implications for the work of a digital repository. A clear positive impact has been the movement towards greater speed in the release and dissemination of data combined with the advantages of cost efficiencies which make the costs of storage and the electronic distribution of data very low, providing a tremendous level of equity in the data sharing process. A clear negative impact is the growth in unrealistic expectations about how quickly data can be prepared and made available to the user community. The technology advances that facilitate the increased ease of delivery has also resulted in amplified expectations over our ability to facilitate the rapid release of newly generated data into the larger world, despite the similar changes in our ability to review, vet and enhance data resources as they arrive at the repository.

This pressure to "rush to publish" has become more common across all the sciences, but in the realm of data management it fails to account for the high degree of effort taken to ensure data is both reliable and safe to use in the research process. Releasing data quickly without a thorough review process greatly increases the risk of error entering the data stream. Human and machine based failures commonly occur as part of data preparation and only the use of highly structured processing pipeline approaches can ensure that these errors will be identified and fixed as part of the formal review of data ingested into the archive repository system. If left unidentified, data errors in shared databases can result in biased analysis, errone-

ous results and incorrect analytic directions. This is particularly the case when users depend upon a growing array of black-box statistical software programs that will provide results, whether the data stream is accurate or error ridden.

In the following sections the paper will first review the way in which data error is perceived in a research data repository; the potential sources of error and bias and what elements the data repository can address within its mission to process, enhance and distribute data for secondary use. The paper will then summarize different approaches that can be taken in the management of large data streams and how different management philosophies can be used to prioritize and control the flow of research information into more realistic packets based upon the scientific value and research demand for types of data.

III. Managing Big Data For Research

A. Evaluating the Risk of Error in Data Specifications

From an archival perspective, survey data are generally considered to be the gold standard for information slated for use in secondary data analysis. This is due to the fact that surveys, when properly administered, provide both high quality and accuracy for the captured outcomes that are meant to represent a defined population or universe. Other data collection approaches, including Randomized Controlled Trials (RCT), Vital Statistics Systems, Censuses and administrative records systems can also have high levels of accuracy, but they face issues less common in survey research. Such data are often captured in stream and may lack internal consistency checks, making harmonization more difficult across time periods. The ability to be generalizable results to the broader population is also a common problem in studies without a representative sampling frame; a particular problem in many RCT studies. Similarly, it may be difficult to share biomedical information and government records due to confidentiality issues or legal restrictions. As a consequence of these and other issues, survey data normally represent the most heavily used information resources within a repository, even if other resources may be larger in size and number of cases. Our experiences with the management of survey data also serve as a useful guidepost for evaluating the potential risks associated with releasing data prematurely, without at least a minimal vetting of the resource for consistency if not accuracy.

In fact, one of the primary concerns associated with adopting a Big Data approach to managing the flow of large electronic data research repositories centers on our ability to effectively monitor the quality of in stream data as it enters the repository and before it leaves for redistribution.

B. The Role of Total Survey Error

In survey research a driving concept underlying quality control is the idea of Total Survey Error (TSE) [12]. This concept tries to account for all possible forms of error that can be introduced into a study during the data generation process. Within a repository system, we seek to further parse this construct into elements that can be addressed during the survey design phase, during the data collection phase, and during the data production phase. Typically, repositories are only actively engaged in the last element, but they often monitor and advise primary data collectors during TSE phases 1 and 2.

In its broadest form, TSE represents the difference between the unknowable true population parameter and the estimate of that parameter derived from the survey itself. This fundamental source of error is composed of two elements: sampling and non-sampling error [13]. Sampling error (Phase 1) results from the unknown variability that exists in any random sample drawn from a defined population. Samples are an estimate of what we think represents the universe of interest, but there are always unmeasured biases that impact the accuracy of this estimate. As survey researchers, we accept there is little that can be done to eliminate the risk of sampling error and it must be assumed that this error will fall within a manageable range if validated methodology is employed. The work addressed in this paper is therefore focused on reducing the other component of TSE, non-sampling error.

C. Identifying Non-Sampling Error (Phase 2)

Non-sampling error represents all the errors that can creep into the survey process, and is often grouped into five general areas: specification error, frame error, nonresponse error, measurement error, and processing error [14]. Specification error occurs when the concept implied by the survey question differs from the concept meant to be measured in the survey. Specification error is often caused by poor communication between the researcher, data analyst, or survey sponsor and the questionnaire designer. Frame error typically results during frame construction when sampling units can be missed or ineligible units inadvertently included. Nonresponse error encompasses both survey nonresponse (the questionnaire is not fielded) and item nonresponse (not all questions are answered completely). Measurement error occurs when the method of obtaining the measurement affects the recorded value, and represents the most studied source of non-sampling error. Finally, processing error refers to errors that arise during the data processing stage, including errors in the editing of the data, data encoding, the assignment of survey weights, and tabulation of the survey data [15].

D. Identifying Non-Sampling Error (Phase 3)

Phase 1 non-sampling errors such as the mistakes in the construction of sampling frames, sample selection and the fielding of questionnaires are generally outside the control of the data archivist as they all occur during the data generation phase. It is in the area of data processing and estimation methods that research archivist has the greatest impact on the reduction of non-sampling error.

IV. Approaches to Big Data Ingestion

In recent years there has been a growing expectation among funding organizations in both the public and private sector that primary data collectors have an obligation to the open, public sharing of data as a condition of support. Recently the US gov-

ernment through a White House directive has required that all federal agencies make plans to provide open and easy access to federally supported data in the very near future [16]. While these changes are seen as a positive expression of the concept that information represents a public good to be shared equitably with the community, it has also resulted in a veritable flood of new studies, surveys and administrative records entering the public domain. In turn, this flow increases the risk of inaccurate or error prone information entering the information super population. Much of these data are managed by data repositories that ingest, process, clean and enhance these files so they are accurate and consistent and introduce as little error as possible into the analysis stream of information.

This growth of openly shared data is expected to increase exponentially in the coming decade, and the effective management of this enormous influx of research data requires that we think of these incoming streams of information within the construct of the Big Data model. While the primary goal will remain the preparation of research data for secondary analysis, these data are expected to come into the repository system faster, with less initial review and less documentation on the part of the depositor. This in turn places a greater burden on the resources of the data repository to perform quality control routines, to update and enhance documentation, often from a static pdf format to a more dynamic XML/DDI presentation scheme.

Additionally, these data will have to be reviewed and prioritized based upon a number of factors, including scientific merit, demand within the research community for type of studies, the quality of the initial deposit, and its relevance to the mission of the repository. More technical issues will also play a role in the priority given to a study, including the physical size of the study, the number of parts, confidentiality risks and the storage format of the study when deposited. All of these factors impact the time and cost associated with making a study both ready for release as an independent item and the integration of the study into the repository super population of data based upon its content and time stamp of collection.

Conversely, the huge influx of Big Data study resources within established repositories has also resulted in higher expectations among users for the rapid release of data that complicates the need for a thorough review and cleaning of files before distribution. This paper reviews, emerging approaches within a research data repository that seek to maintain high quality control while managing Big Data streams as they enter the system.

In the following section the paper reviews different approached to the management of data and its associated metadata (data about data) from the traditional fully processed, best practices model to more recent developments such as open archiving.

A. Current Best Practices in the Preservation of Data

The data curation process, the preparation of data not only for reuse and reanalysis, but also for long term preservation represents an expensive and time consuming activity when fully implemented. Between the extremes of full curation and the destruction of data after its initial use are a wide array of preservation strategies, but for most organizations that actively store and manage data resources there are basic similarities in approach.

Almost all modern repositories employ some variation of the Open Archival Information System (or OAIS) approach [16] as their basic operating model. The OAIS Reference Model [17] illustrates the functions and information flows applicable to a digital archive constructed to maintain safe long-term custody of digital objects. The major functions of the model (Figure 1) include a series of common steps that are employed to greater or lesser degrees by all repository systems:

1) Ingest—the receipt and verification of records as they enter a repository.

2) Archival Storage—the reliable and stable storage of record

3) Data Management—management of records related to the data in a manner that is both accessible and discoverable.

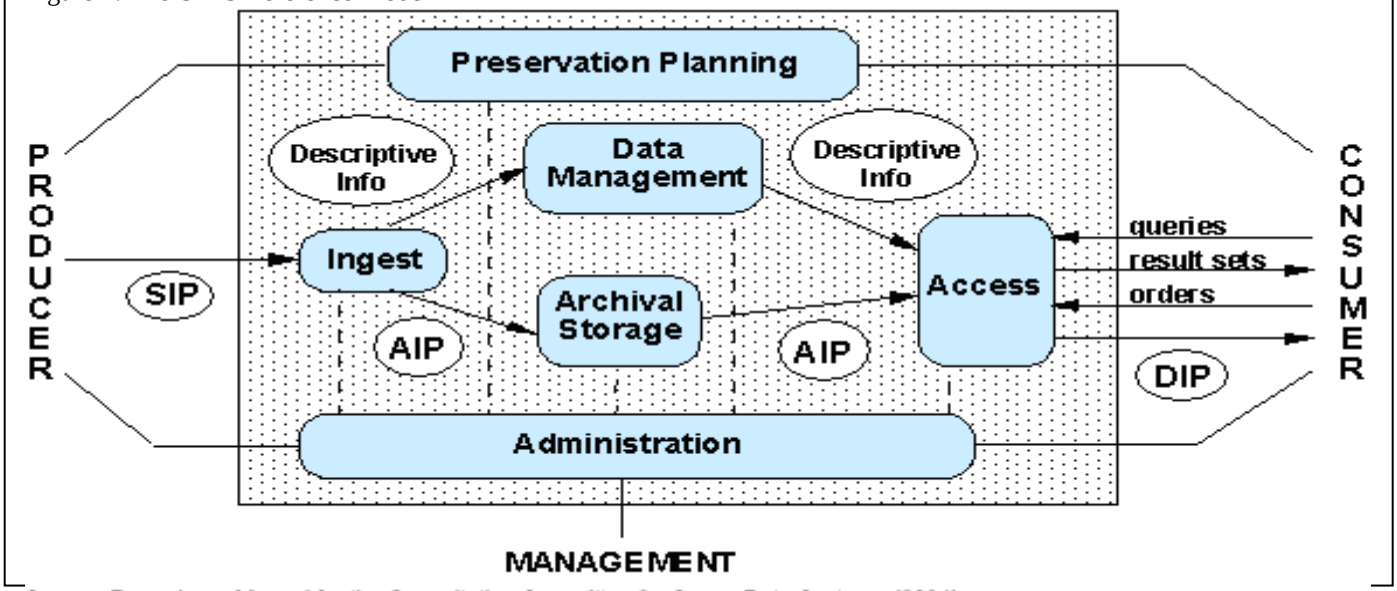
4) Administration—an organizational structure that oversees and manages the individuals who process the data and the relations with the external users of the data.

5) Preservation—the ability to ensure the stability and maintenance of the data and related documentation over time.

6) Access—the ability to provide information and records in response to user requests in a systematic and equitable manner.

The OAIS Reference Model is not a set of strict rules, but rather a series of guidelines that outlines a systematic way to develop and archival structure. By seeking to illustrate the core functions that efficiently maximize information flows without imposing a specific framework OAIS offers a flexible approach

Figure 2: The OAIS Reference Model



Source: Procedures Manual for the Consultative Committee for Space Data Systems (2001)

that is applicable to most repositories regardless of their discipline, mission or research function. One of the goals of the OAIS framework is to allow numerous repositories to exist independent of each other, but with similar underlying structures. Ultimately, this would allow multiple repositories to work together as an information consortium when such a relationship was beneficial and act independently when they had non-complementary goals. This theme is developed further in the next sections as we review different approaches to repository management of large data resources using a generalized OAIS framework.

B. Full Preservation Archiving: The Curation Model

The Inter-university Consortium for Political and Social Research (ICPSR) has operated as a full curation repository since its inception over 50 years ago. The meaning of full curation has changed dramatically across this time period as technology has changed. Up until the late 1990's, curation and preservation were largely restricted to ASCII data stored on 6250 reel tapes and paper documentation. Copies of data were distributed to consortium member through the reproduction of the tapes and the Xeroxing of codebooks with data requests filled through mail delivery. With the growth of the Internet in the 1990s and early 2000s, ICPSR began the process of delivering data through FTP servers and the creation of pdf copies of paper documentation. This process of migration from tape to internet peer to peer delivery would take well over a decade with the last tapes transferred to spinning disc in 2003.

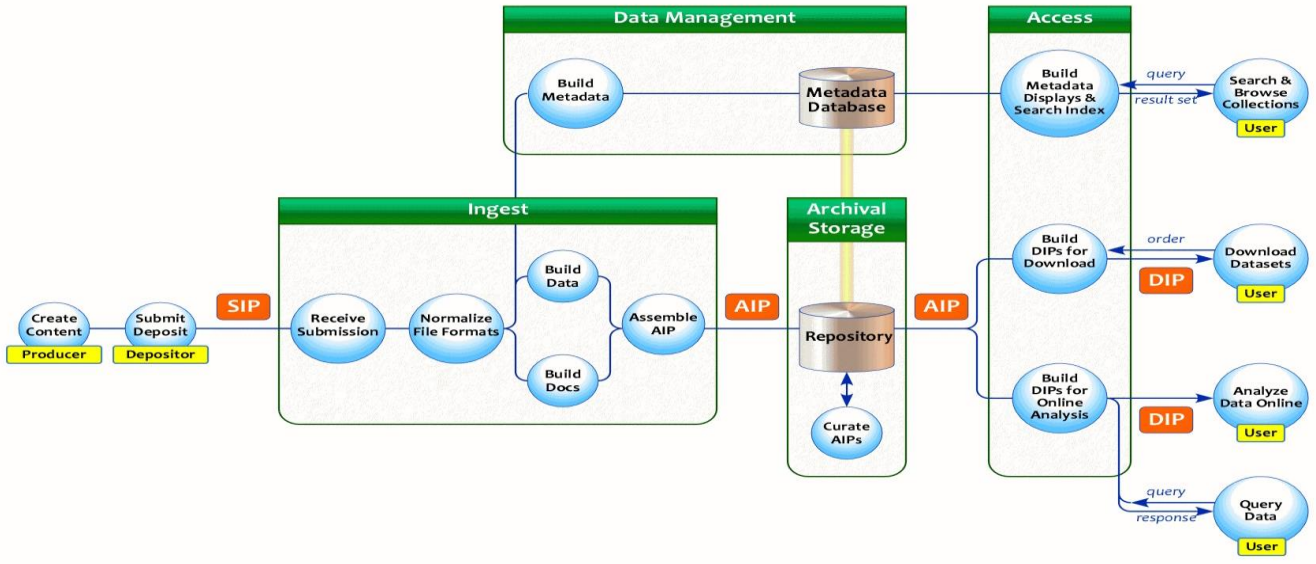
With the movement of all data from tape storage to live disc storage the curation process has become more detailed as it sought to take full advantage of new technology and methods for data preparation. Currently the ICPSR processing pipeline, as illustrated in Figure 2, represents a multistage endeavor from ingestion to distribution [18]. The core of the curation model, however remains consistent with the original data philosophy of the organization, preserving data now so it will always be available in the future. This process is summarized in the Ingestion and Archival Storage phases in Figure 2. Ultimately, ASCII represents the fundamental curation platform. All data, regardless of how it enters the system and regardless of the formats it is distributed in, is preserved as flat ASCII files which remain stable across time and easy to migrate to new storage platforms and read into new or updated software programs. Similarly, documentation is maintained in ASCII format as well, so it will always be readable regardless of how word processing software changes across time. Figures and diagrams are stored as TIFF files and all processing steps are documented and included as part of the curation package. All ICPSR data is curated in a manner that insures that it will be usable and fully accessible in the future as well as meeting current analysis needs of our users.

C. Full Preservation Archiving and Data Distribution Model

While full curation is the foundation of the work done by ICPSR and its Aging Program, the National Archive of Computerized Data on Aging (NACDA), curation alone is not a sustainable approach for a living repository system. It is important for data to be preserved in a manner that makes it available for future users, but it is more important to make these data available for the current cohort of secondary data users. Without a vibrant and user friendly system providing access and distribution system, data maintained in a repository will die a slow death due to lack of use. Consequently, much of the work of ICPSR and NACDA since the early 2000's has been in the introduction of value added products to increase the usability of data in our collections.

The Release and Distribution phases of the processing routines seek to provide as many user friendly options available to the researcher when these data are acquired for reuse from the repository website. The primary value added features are the ability to download specific datasets for use in multiple analysis systems. At present the ingested data is processed and

Figure 2: ICPSR Processing Pipeline Schematic



released in analysis ready files for the SAS, SPSS, STATA and R software systems in addition to raw ASCII files and data definition files. This value added product alone increases our storage needs by more than a factor of four for each study processed as the analysis files themselves will be of various sizes based upon the proprietary compression standard employed by the software producer. ICPSR and NACDA also maintain an online data analysis system that allows users to remotely employ our computing systems at the University of Michigan to perform exploratory analysis of data or to engage in sophisticated multivariate modeling without requiring the data itself to be downloaded to the user's local system.

Documentation is also enhanced as an integral part of the processing pipeline. While the typical documentation, technical files and supporting documents are delivered in pdf format as this is the most common reader used by the research community, underlying the pdf documentation files is XML/DDI markup that allows the base files to be used for a number of additional purposes. Key to this use of the XML\DDI standard it the ability to organize individual dataset and multiple datasets at the discrete variable level. This in turn allows us to build search tools that can identify a single variable of interest within a dataset being queried. More importantly, however, the variable level markup allows us to build more sophisticated search tools that allow internal and external users to seek variables repository wide. A user can search for a specific broad variable such as depression in a study of interest prior to downloading it for use, or they can browse thousands of datasets across the repository simultaneously to identify all studies that have depression variables as part of their content. With the XML/DDI markup at the variable level, the search tools can drill deeper as well, looking for specific kinds of depression measures such as Kessler 6 or the CESD.

As this search can be performed on any variable of interest it provides the user with an invaluable exploratory tool to identify and locate variables of interest across a broad collection of data. The variable markup is an important practical and theoretical move forward in the management of large data collections as it seeks to treat the entire repository collection as a single subset of the super-population [19], [20]. This is the emerging challenge and the exciting opportunity associated with moving repository management into a Big Data world; how do we treat our 10,000 plus studies in a way that treats them as a unified information resource and how do we make this opportunity available to the research community?

Building upon the concept of the Big Data approach as a mechanism for centralizing and unifying information resources, ICPSR also maintains and constantly expands an online bibliography of data related publications [20]. Unlike most online bibliographies, this tool dot not only allows users to identify studies that have addressed specific research topics, it allows the user to see how *specific datasets* have been used for active research resulting in the publication of findings. This bibliography is a dynamic tool, initially integrated into the processing pipeline as part of overall data development, and then maintained as an ongoing process as new publications are identified and associated with the specific study or studies used the publication.

Finally, under the full preservation and distribution model ICPSR provides ongoing support and training in the use of the data maintained in the repository. Direct support is provided to users who have downloaded data from the repository through email, phone and face to face meetings with clients, 5-day to 8-week training course, and a growing collection of training videos, webinars and instructional guides. This process of support serves two fundamental purposes. The obvious one is that this assists the researcher more effectively employ the data for productive research. Secondly, however, it also serves as a key feedback mechanism for the archive, occasionally identifying problems in a particular dataset, but more often providing us with guidance on what features work best for the research community and what steps we might take in the future to provide even more sophisticated value added products and support tools.

D. Full Preservation Archiving and Big Data Challanges

The effort ICPSR has placed into the generation of value added products associated with the data contained in the repository and our more recent efforts in building tools to cross-link and harmonize studies across the collections has resulted in positive benefits for both preservation and for distribution. Providing enhanced data products, analytical tools and user support, the full preservation archiving model used by ICPSR has greatly enhanced the research process and facilitated the more rapid access to and use of data for research development, information exploration and the generation of informed policy. This growth, however, comes at a clear cost when we begin to move more fully into the Big Data environment.

The full preservation, data enhancement and open distribution approach are the most stable and reliable method for handing research data slated for long term use and multiple migrations across platforms as technology and storage platforms change with time. This kind of data management is a classic example of the Long Tail model of digital information management [22]. The costs of a digital repository are almost all concentrated up front; in the costs, resources, and staff time invested in the curation and preparation of the data item for preservation and distribution. Once a data collection is ready to be released for reuse the costs of distribution and storage are almost negligible. Space and bandwidth for moving the data from the repository to the user are relatively low and the data itself can be replicated infinitely with no degrading of the original resource.

Long Tail economics operate very favorably for the user community as the cost of obtaining research data are very small on a per-unit basis once it is prepared for distribution. Unfortunately, these cost efficiencies do not scale across the preservation pipeline. The costs of preparing ingested data for both curation and reuse are largely fixed and do not benefit from the Long Tail economies. Because there is little uniformity in research data deposits it is difficult, if not impossible to automate many aspects of the processing pipeline and

the data curation process remains one largely dependent upon human review and oversight; a costly and time consuming endeavor. All data ingested at ICPSR receives a full confidentiality review regardless of its source of origin, the data is quality checked for errors and inconsistencies, question text is manually extracted from questionnaires and all documentation is reviewed, edited and enhanced. Variable level documentation is created under the XML\DDI guidelines and publications associated with the data are located and associated with the specific data. These and various other tasks are performed on all studies ingested into the repository and all are part of a required set of tasks associated with the processing pipeline.

This level of full curation is essential if long term preservation and the ongoing reliable use of a data resource is to be effectively achieved. It is, however, a time consuming, meticulous and expensive approach to the archival process and few repositories are able to maintain the high standards routinely achieved by ICPSR. Processing time is a particularly challenging aspect of the curation process as it can take anywhere from three months to a year to fully curate a study, depending upon the condition it is deposited in, the number of parts that make up a study collection and the number of times the primary investigator needs to be contacted to address data inconsistencies or concerns. Having a data resource completely resupplied one or more times due to errors is not uncommon and such problems can add months to the processing pipeline before a date resources can be released for reuse and secondary analysis.

From a curation perspective, the amount of time it takes to preserve a study is less central to the process, but most data repositories including ICPSR are not merely in the curation business. Data is curated because it is important to do so, but the data are primarily deposited and then ingested so it can be cleaned, corrected and enhanced for reuse by the research community. With the emergence of the Big Data model as research reality, the expectation of rapid access to large quantities of data has also transitioned into an expectation that new data will also be processed and released with a rapidity that is in direct conflict with the creation of a reliable and stable data resource. Archivally, any amount of data, no matter how large, can be processed over time, but because the costs and time requirements are all front loaded into the data ingestion and preparation phase, little can be done to speed up the pace of the processing pipeline without the sacrifice of data quality and the high risk of error entering a data collection that might be used by hundreds if not thousands of researchers.

While the Big Data explosion has begun, the increasing levels of data deposits processed within the full preservation archiving model employed by ICPSR has not resulted in reductions of data quality, but it has very much resulted in backlogs of new data collections in the processing queue and longer waits for data releases. This problem of balancing the growing quantity of data entering the system against the need for high quality curation is only expected to get worse. ICPSR currently manages petabytes of data and with the expansion of the research curation model to video recording, brain scans, gis and environmental data we are nearing the world of exabyte storage. The exploration of tools and the potential needs of zettabyte management has begun and it seems inevitable that we will enter that storage framework in the coming one to two decades.

This is the challenge the Big Data revolution is bringing to the data repository science and it is one that is almost impossible to solve in a world of fixed and often declining resources earmarked for the preservation of data. Federal agencies, private foundations, and other funders of the research process are increasingly requiring the archiving and open distribution of all research data, but the funds to support such the overwhelming flood of data entering repository systems are not forthcoming. Without a massive infusion of funds to support the archival and curation sciences, the risk of error prone data entering the research stream is largely inevitable as are the risks of confidentiality breaches and compromise of respondent identifies. Like survey research methodology itself, archival preservation and processing pipeline methods are a specialized discipline requiring tools and training not commonly available to the broader research community which may share data with the best of intentions but also with the gravest of consequences if basic best practices are not fulfilled.

At best, the opening efforts to manage the Big Data era for research data within a repository environment will be one of triage; attempts to concentrate efforts and limited resources on producing the most important and essential data collections to the high standards of full preservation archiving. The remainder of the data will need to be dealt with using other approaches which have merits and risks. The following sections of the paper will review the major alternatives to the full preservation approach and evaluate how they might be used to help move the genuine wealth of new information the Big Data era is now bringing to the research community.

E. Open Data Archiving

At the opposite extreme of Full Preservation Archiving Curation models are a variety of low-cost and rapid distribution approaches that allow for open access to research data, but which only partially fulfil the baseline characteristics of the more traditional OAIS based archiving approach. These models fall broadly under the rubric of Open Archiving or Self Archiving models. The approach has clear strengths since the primary requirement for establishing and Open Archival system is the provision of storage space for the data deposits. Open Archiving offers the potential to rapidly ingest and turnover massive amounts of data for distribution in the public domain. It also offers major cost-efficiencies if there is a centralized distribution repository which can represent a weakness for the Self Archiving approach of releasing your own data alone.

Numerous operational Open Archival Repositories exist in various research domains. The social science community has probably made the greatest inroads into testing the full potential of these models. Two examples stand out of how this model can be implemented on a large scale: Data.Gov sponsored by the US Federal government and the Dataverse Project, initiated

by Harvard University and representing an international consortium of Open Archiving repositories.

Data.gov is an open access repository of federal data initiated in 2009 and as of 2015 reported a holdings of 123,456 datasets. Data.gov states it is "a flagship Administration initiative intended to allow the public to easily find, access, understand, and use data that are generated by the Federal government" [23]. While a visionary effort to make federal, state, city and local administrative data more transparent and available to the US user community and the world at large, it has proved to be a difficult vision to maintain.

One of the greatest challenges to maintaining, no less growing the Dat.gov system has been keeping the project funded. In 2011, for example, the Republican Congress implemented a 75 percent cut in overall funds for e-government, efforts which translated to a loss of almost half of the requested funding for the Data.gov project. These funding uncertainties seem to have left Data.gov unable to impose any scalable order or structure into its repository system. Resources are searchable though the available search filter are limited and require a considerable amount of hand review to find relevant items. Of more concern is the lack of standardization in the presentation medium as most data are linked to the local storage at federal, state or city level and often represents a link to an html document that may lead to another website. Metadata is extremely limited and difficult to use in a constructive manner and many links are either broken or out of date. Without the resources to properly manage the repository system Data.gov does not even fully comply with the first three OIAS standards: 1) the ability to ingest data; 2) provision of reliable and stable storage; and 3) the reliable management of data resources, no less the last three stands that support the goal of curation.

The Dataverse Project is an example of a university initiated effort at open archiving and perhaps represents the most successful effort to bring make this approach both user friendly and egalitarian. Dataverse began in 2006 and now claims to have approximately 900 plus independent repositories working under its model which is very consistent with the goals of the OIAS approach for consortium based data sharing. The project itself offers useful search tools and standardized metadata to describe the individual data sets in the collection. Dataverse reports that it has serviced approximately 1.2 million requests for data since 2006 in comparison to the approximately 5 million data requests ICPSR services annually so the Dataverse Project reflects a measurable interest among the user community which may help it maintain the project in the future.

More importantly, the Dataverse Project represents a true Open Archiving model in that any data producer is welcome to deposit their data in the repository by uploading it to the appropriate Dataverse repository. This model again offers a huge capacity to ingest data and to then quickly release for public consumption. The only true limitation on the Dataverse is its storage capacity and its ability to ingest data and process metadata to provide the capacity to search for data types. In spite of the undeniable success of the Dataverse Project as an Open Archival system it shares the same fundamental weaknesses faced by all repositories that focus on redistribution as opposed to curation: sustainability and lack of quality control.

The current Dataverse represents the most recent iteration of open access data sharing generated by a set of capable researchers at Harvard University. Dataverse itself emerged from the ashes of the Harvard Virtual Data Center (VDC) project, which operated during the 1999 to 2006 time period. VDC represented the same philosophy of open archiving and the open sharing of data that is exhibited by the Dataverse Project, but it ultimately failed to attain a sustainable funding model to support its activities and to develop reliable data delivery technology. According to the Dataverse website, Harvard had been experimenting with these kinds of systems since 1987, using early delivery tools such as FTP to move data from one place to another. The point they fail to make is that Dataverse does not represent a system emerging as part of an evolutionary process that builds upon previous successes, but instead it represents a new effort that seeks to overcome the failings that caused the pervious efforts to collapse.

The other issue that is worrisome under the Dataverse project and one that undermines the potential value of the Open Archive model itself, is the persistent lack of quality control over in-stream data being ingested by the repository. Open Archiving is cost effective and fast ways to push data out into the public domain because it requires the data producer to perform the processing pipeline tasks normally performed by the repository under a Full Preservation Archiving Curation model. The data entering an Open Archive repository is deposited as is without serious review and normally without any additional processing or enhancement performed by the repository itself. This is cost effective as it eliminates the high costs that are front loaded into the curation model and do not benefit from the Long Tail economies that make it relatively inexpensive to distribute large quantities of digital material through a website.

While offering huge cost efficiencies in terms of the rapid dissemination of deposited data, the lack of a thorough quality control process prior to release represents a serious concern as virtually anything can be deposited, ingested and distributed through a repository operating under the Open Archive model. These repositories often offer detained guidelines for how the data needs to be prepared, requirements for the removal of identifying variables, and the formats for data files and documentation deposited. Unfortunately, there is little or no way to verify if the depositor has faithfully followed these guidelines and has not inadvertently introduced error or confidentiality disclosive information into the data stream. Though less dangerous to the research process, the lack of quality control can also result in wide variation in data formats, documentation styles and the quality of the data itself. Data preparation for archival and curation purposes is a precise science and archivists routinely repair and revise data submission created by well-meaning researchers who are not properly trained in the best practices of archival preparation.

In sum, the Open Archiving model has attractions and clear cost efficiencies, but it does not provide a standardized data product and it has few, if any safeguards against the risks of inadvertent identity disclosure, the ingestion of poor quality data, and the risks of error prone data being introduced into the research community and then naively used for research purposes. Both the lack of standardization and quality control make

the Open Archive approach inappropriate for successful management of Big Data information streams.

F. Virtual Data Archiving

Standing in a middle ground between Full Preservation and Open Archiving models is the concept of Virtual Archiving. Under a Virtual Archiving model the repository captures or creates detailed metadata describing the data resource and makes the resource discoverable through a repository search engine. Like Open Archiving, the Virtual Archiving model can offer low cost and rapid turnover of data resources depending upon how it is handled by the repository itself. In the most minimal approach, the repository can create a simple metadata record that provides a summary description of the data element and then provides a stable link to the external website that would supply the data. This approach was used years ago by ICPSR to establish what it called a Union Catalogue of external data resources though new approaches has since been introduced. A slightly more sophisticated approach is to include more detailed information and separate links to specific documentation and data files maintained on the external website, an approach used by many contributors to the Data.give website.

An emerging model initiated and promoted by ICPSR is the creation of a complete metadata record for an eternally distributed dataset or collection. Under this model, not only is a detailed description of the data provided and fully discoverable within the repository search systems, the study documentation is also captured. The documentation is processed though the ICPSR pipeline without associated data so a XML\DDI documentation record can be created for the file. The creation of the XML\DDI documentation allows ICPSR to create a variable level codebook that then allows interested researchers to search the external data contents and identify specific variables of interest before moving on the external website to request the data. The creation of variable level information also allows the virtually managed dataset to be introduced into the ICPSR variable level database which allows the information from these externally supported data collections to be compared across the ICPSR repository collections. This feature adds additional levels of discoverability to the virtual data collection as board search for variables will return results not only on data physically present in the repository but also for studies identified but not physically archived at ICPSR.

The Virtual Archive model shares many of the benefits and weaknesses associated with the Open Archive model. Virtual Archiving is more cost effective that the Full Preservation Archiving approach as the data associated with the metadata records do not have to be processed by the repository and the release of the discoverable metadata record can be facilitated quickly as no data review is required. Archiving data through a virtual approach, however, does make the production of standard data products impossible across virtual acquisitions. The documentation can be standardized and released in line with current best practices, but associated data are not physically touched under this approach, only referenced and then the user is redirected to the origin site where the physical data is maintained. The actual data distribution may produce and distribute their data under any form they find appropriate regardless of data production specifications required for products fully ingested by ICPSR.

Similarly, the quality control of the data itself remains external to the virtual archive that captures the associated metadata. While the repository archive can enhance and correct metadata and documentation files captured under the virtual archiving process, it cannot repair or change any aspect of the analysis data and the distributor of origin is fully responsible for processing the data files to whatever standard they use for preparing data for reuse and secondary analysis.

While far from a perfect solution for uniform curation of data products when compared to the Full Preservation Archiving model, the Virtual Archiving approach is preferable to the Open Archive model for a number of reasons. First, at least some elements of the data collection can be managed and enhanced to best practices through the creation of XML\DDI documentation to be associated with the metadata files. This provides a higher and more sophisticated level of discoverability for the external data and it allows many elements of external data collections to be integrated into the overall repository, introducing these external variables into the searchable index of variables maintained by ICPSR and expanding the known universe of research information beyond what is physically maintained within our collections.

The use of a Virtual Archiving model also have two additional benefits not found in the Open Archive approach: acquisition control and deniability. One of the strongest concerns with a true Open Access Archive is the inability to refuse the submission of data not deemed appropriate to the repository mission without ongoing oversight and review of deposit is entering the ingestion stream. The cost efficiencies of an Open Archival model are largely dependent upon all depositors following guidelines on what kind of data is acceptable and how this data should be prepared for deposit. Once data is seen as potential risk due to non-compliance with the Open Archiving philosophy, then either the repository accepts this risk or it institutes a review process which degrades the cost efficiencies of the model. Under a Virtual Archiving model, the repository actively selects which data to capture remotely and is under no obligation to accept external deposits without review. In a similar manner, the deniability aspect allow the repository to place a layer of protection between their repository and the risk of ingesting error prone of poorly processed data. Under an Open Archive model, the data, whether good or bad, physically resides within the repository of record and regardless of written warnings or disclaimers associated with the data released under an Open Archive, the ingestion of unvetted data carries a clear risk that the reputation of the repository could be damaged if disclosive or error prone data resources are released through their system.

V. CONCLUSIONS

The traditional world of fully processed and curated data resources is changing in the face of the realities of a Big Data world where massive amount of data are released for reuse and reanalysis by the information community. The archival process has emphasized the slow and thorough review of a limited number of high impact studies that are used heavily by a large cohort of researchers across time. With the shift to a Big Data mentality, researchers increasingly want more data, and they want it quickly. "More and quicker" do not lend themselves to the generation of reliable and error free data resources, as the inflow of research data to repositories is rarely consistent and no one single standard exists for how data should be prepared for secondary use by the primary investigator. Normally, the archival specialists would ingest data in a variety of forms and formats, decompose them to their base parts and then build a data product that would be consistent in structure and accessibility to all other data collections within the repository. This is process that is high both in cost and in human intervention even with technological advances in storage and data management tools.

Alternative approaches to managing large quantities of disparate research data such as Open Archiving and Virtual Archiving offer some relief as these approaches can offer rapid turnover capacities at a relatively low cost. This is done, however, at the expense of quality control and detailed review of the data ingested into such systems. In order to address the data needs and requirements of the upcoming generation of Big Data, these approaches, however, will have to be grafted onto the older curation approach. Through the blending of archival approaches and the careful creation of a triage system to determine which pathway specific data need to directed into as part of the preservation and distribution process it is likely that the anticipated flood of data can be managed and preserved for the long term. Some data can be fully processed, some lightly process and some simply preserved in a stable archival format for future use.

In the long run all data can be preserved, but the existing expectation that Full Preservation Archiving repositories such as ICPSR can fully curate all research data that comes into the archive is misplaced. In light of declining resources and the ongoing demand for the creation of value added products to enhance and facilitate research decisions will increasingly need to be made as to which data can receive these services and which data will need to accept a lesser level of archival investment. These choice will be made, but like the curation of data itself, they should not be made in haste.

REFERENCES

[1] Rider, Fremont 1944. The Scholar and the Future of the Research Library. A Problem and Its Solution. Hadham Press (1944)

[2] B. A. Marron and P. A. D. de Maine. Automatic data compression. Communications of the ACM CACM. Volume 10 Issue 11, Nov. 1967, Pages 711-715The Total Survey Error Approach: A Guide to the New Science of Survey Research, University of Chicago Press ISBN 0-226-89128-3

[3] Becker, Hal B., "Can users really absorb data at today's rates?, Tomorrow's ?". Data communications, July 1986, PP 177-193.

[4] Demming, Peter. "Saving all the bits" American Scientist 78, 5 (September-October 1990), 402-405.

[5] Morris, R.J.T. and Truskowski, B.J. The evolution of storage systems. IBM Systems Journal. Volume:42 Issue:2.

[6] Diebold, F.X. (2003). "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting". in M. Dewatripont, L.P. Hansen and S.Turnovsky (Eds.), Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society. Cambridge: Cambridge University Press, 115-122.

[7] Laney, Doug (2001). "3D Data Management: Controlling Data Volume, Velocity, and Variety.". Research Note for the Meta Group Application Delivery Strategies. Feburary 6, 2001.

[8] Hannay, Timo. (20140. Science's Big Data Problem. Digital Science. 08.26.14. URL: www.wired.com/2014/08/sciences-big-data-problem/

[9] Boyda, Danah and Kate Crawford. (2012) "Critcal Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon". Information, Communication & Society. Special Issue: A decade in Internet time: the dynamics of the Internet and society. Volume 15, Issue 5, 2012. DOI:10.1080/1369118X.2012.678878

[10] Fairfield, J., & Shtein, H. (2014). Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. Journal of Mass Media Ethics, 29(1), 38-51.

[11] The SAS Institute. (2013) Five big data challenges and how to overcome them with visual analytics. URL www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[12] Assael, Henry; Keon, Jhn (1982) "Nonsampling vs. Sampling Errors in Survey Research", The Journal of Marketing 46 (2), 114–123 JSTOR 3203346

[13] Biemer, P.; Lyberg, L. (2003). Introduction to Survey Quality. John Wiley & Sons, Inc. ISBN 0-471-19375-5

[14] Groves, R.; Fowler, F.; Couper, M.; Lepkowski, J.; Singer, E.; Tourangeau, R. (2009). Survey Methodology (2nd Edition). John Wiley & Sons, Inc. ISBN 0-470-46546-8

[15] White House Office of Science and Technology Policy. "Increasing Access to the Results of Federally Funded Scientific Research". Memo February 22, 2013 URL: www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[16] Consultative Committee for Space Data Systems. (2012). Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS Secretariat. URL: public.ccsds.org/publications/archive/650x0m2.pdf oung, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[17] Sawyer, Don. 2000. The Open Archival Information System and the NSSDC. NSSDC News. December 2000 Issue. URL: nssdc.gsfc.

[18] Inter-university Consortium for Pilitical and Social Research (ICPSR). (2012). ICPSR's Guide to Social Science Data Preparation and Archiving, 5th Edition. URL: www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html

[19] Little, R. J. (1982). Models for nonresponse in sample surveys. Journal of the American Statistical Association, 77(378), 237-250.

[20] Hartley, H. O., & Sielken Jr, R. L. (1975). A" super-population viewpoint" for finite population sampling. Biometrics, 411-422.

[21] Inter-university Consortium for Pilitical and Social Research (ICPSR). (2014). ICPSR Bibliography of Data-related Literature.URL: www.icpsr.umich.edu/icpsrweb/content/ICPSR/citations/methodology.html. Last updated May, 2014.

[22] Anderson, Chris. (2004). The Long Tail. WIRED Magizine. Issue 12.10 - October 2004. URL: archive.wired.com/wired/archive/12.10/tail.html

[23] Office of E-Government and IT, Office of Management and Budget. (2009). Data.gov Concept of Operations (Draft). URL: www.ideascale.com/userimages/sub-1/736312/ConOpsFinal.pdf