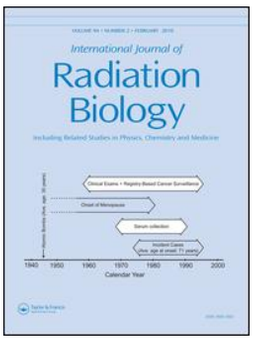


International Journal of Radiation Biology

ISSN: 0955-3002 (Print) 1362-3095 (Online) Journal homepage: https://www.tandfonline.com/loi/irab20

Big data in radiation biology and epidemiology; an overview of the historical and contemporary landscape of data and biomaterial archives

Paul N Schofield, Ulrike Kulka, Soile Tapio & Bernd Grosche

To cite this article: Paul N Schofield, Ulrike Kulka, Soile Tapio & Bernd Grosche (2019): Big data in radiation biology and epidemiology; an overview of the historical and contemporary landscape of data and biomaterial archives, International Journal of Radiation Biology, DOI: 10.1080/09553002.2019.1589026

To link to this article: https://doi.org/10.1080/09553002.2019.1589026

Big data in radiation biology and epidemiology; an overview of the historical and contemporary landscape of data and biomaterial archives

Paul N Schofield[1], Ulrike Kulka[2], Soile Tapio[3], and Bernd Grosche[4]

[1] *Dept of Physiology Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK*
[2] *Bundesamt fuer Strahlenschutz,Ingolstädter Landstraße 1, Neuherberg 85764, Germany*
[3] *Helmholtz Zentrum Muenchen, German Research Center for Environmental Health GmbH, Institute of Radiation Biology, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany*
[4] *Retired; formerly Bundesamt fuer Strahlenschutz, Germany*

Abstract

Purpose: Over the past 60 years a great number of very large datasets have been generated from the experimental exposure of animals to external radiation and internal contamination. This accumulation of "big data" has been matched by increasingly large epidemiological studies from accidental and occupational radiation exposure, and from plants, humans and other animals affected by environmental contamination. We review the creation, sustainability and reuse of this legacy data, and discuss the importance of Open data and biomaterial archives for contemporary radiobiological sciences, radioecology and epidemiology.

Conclusions: We find evidence for the ongoing utility of legacy datasets and biological materials, but that the availability of these resources depends on uncoordinated, often institutional, initiatives to curate and archive them. The importance of open data from contemporary experiments and studies is also very clear, and yet there are few stable platforms for their preservation, sharing, and reuse. We discuss the development of the ERA and STORE data sharing platforms for the scientific community, and their contribution to FAIR sharing of data. The contribution of funding agency and journal policies to the support of data sharing is critical for the maximum utilisation and reproducibility of publicly funded research, but this needs to be matched by training in data management and cultural changes in the attitudes of investigators to ensure the sustainability of the data and biomaterial commons.

Keywords: Big data, database, archive, radiobiology, epidemiology, open data, biorepository, data sharing.

Introduction

The management, availability and sustainability of scientific data has become a critical issue for the biomedical sciences over the last decade, with increasing political, scientific and social concern about the issues of data sharing, accountability and reproducibility in the biological sciences (Baker 2016, Nature Editorial 2017). This has drawn attention sharply to the accessibility of primary and even derivative data, not only those associated with publications but in many cases large datasets that have neither been used for publication nor been put into the public domain. Therefore such proprietary datasets, human clinical trial and epidemiological data, and legacy data are at risk of being lost. It is the aim of this commentary to look at the historical development of large key datasets in the radiobiology and to review efforts to provide open data repositories across all domains of investigation with the aim of sharing data across the community.

It is important to define what we mean by data archives and datasets. Much of the data collected into historical archives – i.e. those whose collection has ceased - to all intents and purposes closed legacy datasets – are derived from very large experiments which are often loosely based on hypothesis testing, and not designed to test specific biological or physical mechanisms; i.e. a wide range of data collected in order to inform a broad question. This includes lifespan studies, cancer studies and those with broadly defined endpoints. Such archives include the very large human radiation exposure datasets, some of which are still collecting data – large-scale epidemiological datasets, and the results of extremely large-scale animal exposure experiments. In all cases we can legitimately describe these as “Big data” – some of these were possibly the largest and most complex data collection exercises in the biological sciences conducted to the date they were completed. Examples are the important epidemiological study of the Japanese atomic bomb survivors (Ozasa et al. 2018) and the large Million worker study which includes worker cohorts from U.S. Department of Energy (DoE) Manhattan Project facilities, nuclear power plants, industrial radiographers, U.S. Department of Defense (DoD) nuclear weapons test participants, and medical technicians and physicians (Boice et al. 2018). Data complexity characterises big data as much as volume, and large complex datasets are

at the same time more difficult to manage and more potentially fruitful in analysis. The characteristic of this type of data, and indicative of its ongoing value, is that it is possible to reanalyse, recode, integrate and aggregate data, and to reinterpret it according to changing scientific paradigms. The size of some datasets might lend itself to machine learning approaches, for example to generate classifiers, but trained or deep learning methodologies such as deep neural networks have not yet to our knowledge been applied in radiation epidemiology or animal irradiation experiments. However machine learning is being applied to discovery of radiation specific transcriptional signals for example (Zhao et al. 2018) and increasingly in radiotherapy and medical physics (Sahiner et al. 2019). Successful application of support vector machines in determining the directionality of aerial radiation dispersal (Yoshikane and Yoshimura 2018) provides a model for retrospective studies where sufficient data is available.

Some of these archives have associated with them physical specimens, blood, tissue, or histopathological slides. We discuss below several of these physical resources that include significant patient or animal data as part of their structure. One such example would be the Chernobyl thyroid tissue bank (Thomas 2012). Another class of archive is derived from literature curation or integration of primary data with literature derived data. As they are subjected to expert manual curation these resources can be very valuable.

There have been few attempts to systematically aggregate data, pointers to data or physical resources in radiation biology, but such aggregations can legitimately be termed data archives. Some of these are closed and constitute a stable resource for legacy data, such as the ERA database (see below). Others such as the STORE database (see below) have been developed to act as an open repository for either legacy or ongoing studies as a mechanism for preserving data and information about bioresources, and disseminating it to the community.

Firstly, we consider first legacy archives and projects that are effectively closed to further data accretion together with efforts to make them accessible and useable. Secondly, we discuss archives of long-term experiments and epidemiological datasets that are still accumulating data, and consider archives of physical resources such as

organisms, tissues, blood and non-biotic material. A summary of the resources we discuss can be found in Table 1. Finally, we discuss the archiving and dissemination of data from active studies and that deposited as part of currently funded work and publications, together with a consideration of the current emphasis on open data and the issues surrounding compliance with open data mandates in the community. Our intention was to include datasets and archives from various areas of radiation research to address the point we want to make about the importance of archiving and data sharing. We are aware that the list of datasets and archives mentioned in this review cannot be complete and that more databases exist which were not included in the current work.

Legacy data archives

Beginning in the late 1890s with the discovery of X rays and then Radium (Sekiya and Yamasaki 2016) early animal and human exposures were often accidental and sporadic with a small number of individuals involved. In the early part of the 20th century, with more radionuclides becoming available, small quantities were widely used in patent health products, particularly Radium, for example in the radium containing drink, Radiothor, which contained 74kBq of a mixture of radium 226 and radium 228 in each bottle (Macklis 1990). These patent medicines and other products were considered to confer health benefits until, after some notable deaths, radiopharmaceuticals were brought under regulation in the early 1930s, shortly after the time when the mutagenic action of radiation exposure was definitively established (Muller 1927). One of the first large-scale data collection exercises concerned internal occupational exposure of US radium dial painters (Fry 1998). Long-term studies of these workers traced 1322 women first employed between 1913 and 1929, and 1403 women first employed between 1930 and 1949. Follow-ups and analysis have continued up to the late 1990s demonstrating the importance of long-term studies and data sustainability along with the analysis of data not previously envisaged when it was first collected. As radiation began to be used clinically and its effects were beginning to be appreciated, X-irradiation was widely employed, and in some cases on a very large scale. For example, cranial X-irradiation was used in the treatment of the fungal scalp disease *Tinea capitis* over the period 1948-1960 which was the

subject of a very large follow-up study beginning in 1968 (Sadetzki et al. 2005) whose data remains available.

Legacy data from human exposure

The stimulus for large-scale animal and human experimentation with radiation exposure was a consequence largely of the United States nuclear weapons programme and the subsequent release of the first nuclear weapons over Hiroshima and Nagasaki. It is no coincidence that the foundation of the International Journal of Radiation Biology in 1959 coincided with the surge of interest in the acute and long-term effects of radiation, and to a great extent the history of big data in radiobiology is parallel to the history of this journal.

In 1946, following Congressional hearings, the US Atomic Energy Commission was established and shortly afterwards, in 1947, its chairman David Lilienthal commissioned a Medical Board of Review, to report on the agency's biomedical program (Hewlett et al. 1990). The board strongly recommended a broad research and training program:

> "..both urgent and extensive." The need is urgent because of the extraordinary danger of exposing living creatures to radioactivity. It is urgent because effective defensive measures (in the military sense) against radiant energy are not yet known."[1]

There was increasing public concern about the effects of nuclear fallout and especially after the leakage of data concerning the impact on bystanding observers and the local population following the testing of US nuclear weapons over Bikini atoll in the Marshall islands in the early 1950s. Operation Crossroads and Operation Castle Bravo generated serious concern about the danger of irradiation and contamination, particularly in the light of the first alarming analyses of the Japanese A bomb

[1] "Report of the Board of Review," 20 June 1947, attached to letter from David Lilienthal, Chairman, AEC, to Dr. Robert F. Loeb, Chairman, AEC Medical Board of Review, 27 June 1947 ("At the conclusion of the deliberations . . .") (Advisory Committee on Human Radiation Experiments (ACHRE) No. DOE-051094-A-191), 3-4 avalable at available at <http://www.gwu.edu/~nsarchiv/radiation/> and <http://www.eh.doe.gov/ohre/roadmap/achre/report.html>.

survivors (discussed below) in the immediate aftermath of Hiroshima and Nagasaki. While it is not the aim of this commentary to unpick the political and economic events which led up to the first large-scale animal testing of radiation exposure, the motivation for these studies, the data from which has still not been exhaustively analysed, constitute some of the largest datasets in radiation science.

Following the initial studies on the survivors of Hiroshima and Nagasaki, a significant number of experiments were carried out on human subjects between 1940s and 1970s which came to light in the early 1990s (McCally et al. 1994, Stone 1993). Primary data are scattered through US agencies and universities and have not so far been made public to our knowledge, though these datasets would certainly benefit current research. Below we consider the collection of large-scale data on the Japanese A bomb survivors and other human exposures, both occupational and accidental, before moving to a consideration of the major animal exposure experiments starting in the late 1950s.

Hiroshima and Nagasaki survivors; the LSS study

Following the dropping of two atomic bombs over Nagasaki and Hiroshima in 1945, the Atomic Bomb Casualty Commission (ABCC), now the Radiation Effects Research Foundation (RERF), was set up in 1946 to monitor the health of the survivors. By the end of 1945 more than 200,000 had died of the combined effects of physical injures, acute radiation sickness and late effects. By 1950 there was also concern about gonadal doses and germ line mutation. The lifespan study cohort (LSS) was established in 1958 comprising standardised data on 120,321 individuals, including co-resident but unexposed controls (Ozasa et al. 2018). Further cohorts have also been established (Ozasa 2016): the adult health study (AHS) aimed at gathering morbidity data for disease additional to cancer, and the In Utero programme focussed on 3,268 individuals exposed in utero. A third study examines the heritable impact of exposure, the 'F_1' study, which aims at elucidating the impact of radiation exposure on the germline. Summary data for all these cohorts are available, but access to detailed individual level data requires RERF approval.

Occupational and accidental exposure in the Soviet Union

Human exposure data for the period starting in the 1940s up until the early 1980s are available from the Mayak plant in the Southern Urals in Russia, derived from close monitoring of workers in Mayak where from, starting from the late 1940s, highly enriched uranium, tritium and plutonium was produced for Russian nuclear weapons. Occupational exposure and accidents were recorded between 1948 and 1982, with more than 30% of workers estimated to have been exposed over the working lifetime to more than 1Gy of mainly external γ doses, the average internal ^{239}Pu contamination being 2.19 ± 0.15 kBq (Azizova et al. 2008). The data consist of ICD9-coded medical records, doses, cause of death, work history and demographic information on 12, 585 workers, and are augmented by biological samples, both from blood and autopsy. Tissue collections and data from this resource have been used with considerable impact for example on studies of cardiac exposure (Azimzadeh et al. 2017).

Distinct from the Mayak cohorts are the studies on the Techa river where over nearly a decade, starting from 1949 the Mayak plant discharged liquid radioactive waste (7.6e6 m^3) into the river, thereby polluting large areas of the surrounding region and exposing the surrounding population to long term internal contamination. Data have been collected from this area since the early 1960s including demographic and clinical information from approximately 29,000 inhabitants. These data contain information on sex, cause of death, period of exposure and estimates of dose. The Techa river database is one of the few containing information about protracted environmental radiation exposures in a general population (Krestinina et al. 2005).

Semipalatinsk Nuclear Test Site

From 1949 to 1989 nuclear weapons testing was conducted by the former Soviet Union at the Semipalatinsk Nuclear Test Site, Kazakhstan, including 111 atmospheric or near-ground tests between 1949 and 1962. Four nuclear weapons tests, conducted from 1949 to 1956, resulted in non-negligible radiation exposures to the public, corresponding up to approximately 300 mGy external dose. The population living around the test site is one of the largest human cohorts exposed to radiation from nuclear weapons tests. As a follow-up of research that started in the 1960s, a registry

that contains information on more than 300,000 individuals residing in the areas neighbouring the test site was established. The registry contains relevant information about those who lived at the time of the testing as well as about their children and grandchildren, including to some extent biological material (Apsalikov et al. 2019). To date, only a few studies have been conducted which were either completely (Grosche et al. 2011) or partially (Land et al. 2008) based on the information from a precursor of the registry. The registry can now be used for future studies, and detailed information on a data set for a three-generation study is already included in STORE (http://dx.doi.org/doi:10.20348/STOREDB/1091).

Wismut uranium miners study

The WISMUT study contains data on approximately 59 000 male uranium miners, first employed between 1946 and 1989, at the Wismut Company in Germany. It contains demographic, cancer and other mortality data. It is the largest single study on the health risks of occupational exposure to ionising radiation and inhalation of radioactive radionuclides in uranium mining (Kreuzer et al. 2010). The data can be accessed through the STORE database (http://dx.doi.org/doi:10.20348/STOREDB/1036) (see below).

German Thorotrast study

The thorium-containing radioactive contrast agent Thorotrast® was used from 1929 until the1950s as a contrast agent in angiography and arteriography. The thorium in Thorotrast persists throughout the lifetime of the exposed patients who consequently are exposed to a lifetime's chronic internal exposure. Several cohort studies were initiated, notably in Germany, and the German Thorotrast study cohort was established retrospectively in 1968 with a follow-up until 2004. The study comprises 2326 Thorotrast patients and 1890 patients of a matched control group. The dataset contains demographic, dosimetric, morbidity and mortality data and can be obtained on application through the STORE database (http://dx.doi.org/doi:10.20348/STOREDB/1016) (Grosche et al. 2016)).

Japanese Thorotrast study

Parallel to the above, a study of 436 Thorotrast-exposed patients was also carried out in Japan and both patient data and material are available (http://www2.idac.tohoku.ac.jp/misc/thorotrast/index%20english.html). Data includes estimates of thorium amount deposited and cumulative dose in major organs, and confirmed pathological diagnosis (Fukumoto, 2014).

Kyshtym, Chernobyl and Fukushima

Six nuclear accidents have occurred in the past, Kyshtym (1957), Windscale Piles (1957), Three Mile Island (1979), Chernobyl (1986), Tokaimura (1999) and Fukushima (2011).
In the accident at the Mayak plant on 29th September 1957 (the **"**Kyshtym Accident'**)** (Akleyev et al. 2017) 20 MCi **(**740 PBq**)** of radionuclides were released from a chemical explosion on the site. The subsequent spread of contamination was monitored, and the exposed population enrolled into the database of the URCRM, which contains the results of long-term dosimetric monitoring and medical follow-up of the population. The cohort contains around 21, 000 individuals being, along with the Techa river cohorts, one of the largest prospective datasets available from accidental contamination of civilian populations.

The Chernobyl accident in 1986 affected the populations in Ukraine and Russia but mainly Belarus. In addition to the affected general population around 600,000 workers were involved in the cleanup operation. The cleanup workers were mainly exposed to γ radiation with an estimated mean dose ranging from 20 to 185mGy. There have been several overlapping studies performed on these populations with endpoints including thyroid cancer, leukaemia and lymphoma. Both closed and continuing studies being subject to intensive analysis, reviewed comprehensively by Cardis et al. (Cardis and Hatch 2011, Hatch and Cardis 2017).

The Fukushima Daichi Nuclear power plant incident in 2011, following the Tōhoku earthquake and tsunami, involved a core melt-through damaging three reactor cores followed by hydrogen explosions. As with Chernobyl, both the local population and

emergency workers were exposed to external mainly γ radiation and internal contamination with a maximum external dose to emergency workers of around 700mSv and residents around 25mSv (Hasegawa et al. 2015). Large-scale health surveys of the TEPCO emergency workers are being established by RERF – the NEW study (Kitamura et al. 2018), with around 5000 workers having been recruited to date. The Fukushima Health Management Survey of Fukushima residents (Ishikawa et al. 2015) was created by the Fukushima prefecture and contains dose estimated for individuals, based on their movements during the accident, and overall health assessment, thyroid ultrasound examination, mental health and lifestyle survey, and a pregnancy and birth survey. Emerging data from the epidemiological studies suggests that a very significant measure of morbidity has its origins in psychological aspects of displacement, or fear of radiation and social issues, and it will be interesting to see how future analysis of these psychosocial datasets feeds into future disaster planning and mitigation strategies.

Comprehensive Epidemiologic Data Resource (CEDR Database); U.S. Department of Energy

The CEDR is the U.S. Department of Energy (DOE) electronic database that contains de-identified data on health studies of DOE contract workers and environmental studies of areas surrounding DOE facilities. The resource currently contains 76 studies of over 1 million workers at 31 DOE sites. Much of the data is from epidemiological studies at US nuclear facilities and provides access to individual level data in many cases, with primary raw and derived datasets. A complete description of the data and the resource can be found on (https://apps.orau.gov/cedr/).

Additional human datasets

An excellent review listing the major human epidemiological datasets available – with a focus on cardiovascular diseases – was published recently (Kreuzer et al. 2015) although access to these datasets is largely on a discretionary basis where there are issues of data consent and local personal data legislation. Notably included in these large datasets are the International Nuclear Workers Study (INWORKS) (Hamra et al. 2016), an integrated study of more than 380, 000 nuclear workers in three countries

(USA, UK, and France), and that of the UK nuclear workers, UK NRRW (Haylock et al. 2018) which is partially proprietary. A large cohort of 948,174 children (with follow-up data) exposed to ionizing radiation by CT scans was set up as a joined effort of nine European countries (Bernier et al. 2018). As with the INWORKS study, these data are proprietary and held at IARC, Lyon.
A more comprehensive description and discussion of human datasets has been published recently (Zander et al. 2019).

Large-scale animal experiments

In the early 1950s there were significant concerns about the scientific utility and ethics of radiation exposure experiments on humans. Sheilds Warren, the Chair of the AEC reported in 1950[2] (cited in (Faden 1996):

> "We have learned enough from animals and from humans at Hiroshima and Nagasaki to be quite certain that there are extraordinary variables in this picture. There are species variables, genetics variables within species, variations in condition of the individual within that species." The danger of failing to provide data had to be weighed against the danger of providing misleading data: "It might be almost more dangerous or misleading to give an artificial accuracy to an answer that is of necessity an answer that spreads over a broad range in light of these variables."

In 1951, following the Operation Greenhouse hydrogen bomb tests on Enewetak, 4000 mice exposed to radiation from the blast were taken to Oak Ridge and received by Jacob for long-term study (National Academy of Engineering 1984). This was the beginning of a very large series of non-human mammal internal and external exposure experiments. From Warren again:

> "Jacob was the recipient of large numbers of mice, survivors from a Pacific nuclear test, placed with various degrees of shielding along radii from the point of explosion. He had the foresight to follow these animals to the time of their natural death. As a result of these studies, much new information was developed about the late effects of radiation, about biological dosimetry, and about the similarity of certain radiation effects to those of aging."

Between 1952 and 1992 more than 200 large-scale experiments were conducted on non-human animals, mainly mice and beagles, in the USA, Europe and Japan. For

[2] Warren, transcript, Advisory Committee for Biology and Medicine, transcript (partial) of proceedings of 10 November 1950 (ACHRE No. DOE-012795-C-1) of 10 November 1950, 13.

example, at Argonne National Laboratory (ANL) 700 beagles and 50, 000 mice were used in experiments between the late 1960s and early 1990s as excellently reviewed by Haley et al. (Haley et al. 2011). This included the JANUS studies on whole body γ and neutron radiation of inbred strains of *Mus musculus* but also *Peromyscus sp*. funded by the now Department of Energy, which emerged from the AEC. These were generally lifespan studies and involved detailed cross-sectional, longitudinal and terminal pathological investigation over a wide range of irradiation doses, dose rates, quality and timing.

The Argonne beagle dog experiments, carried out at Argonne National Laboratory, the Pacific National lab, UC Davis, and University of Utah from 1952 to 1991 and supported by grants from the Atomic Energy Commission, investigated the effects of ^{60}Co radiation on nearly 5000 beagle dogs. In addition, internal contamination with radium, Pu, Cf, and ^{90}Sr, was investigated – the latter considered an important component of nuclear fallout. Types of exposures ranged from external radiation to inhalation and using acute, chronic and fractionated doses.

Taken together these large-scale mammalian studies form the basis of much of our knowledge concerning the acute and chronic long-term effects of external and internal radiation, and constitute a huge data resource. While some of the data, or at least data analyses have been published, by the 1980s it was clear that the primary data from these experiments were in danger of being lost. Given the high estimated cost of $2bn, at current costs, needed to repeat these experiments even if the necessary infrastructures were still available, it became apparent in the 1980s that it was desirable to salvage this legacy data and put it into the public domain for further use and analysis. Consequently, the data from the Argonne Janus mouse studies carried out between 1969 and 1992, including around 50, 000 mice, was curated (Wang et al. 2010) and is now housed in the Northwestern University Radiation archives (NURA) along with beagle data from ANL which includes data from thousands of dogs in mainly lifespan studies. Both datasets have associated tissues, also preserved at NURA (Haley et al., 2011) and are freely available. The data and tissues archived at NURA have been used for new analyses, for example the effects of radioprotective agents (Paunesku et al. 2008), interspecies sensitivity (Liu et al. 2013) and gender effects (Haley et al. 2011).

The European Radiobiological Archive (ERA)

In the mid-1980s, the European Late Effects Project Group (EULEP) embarked on an initiative to collect and collate data covering all available information on European long-term radiobiological animal studies. The Office of Biological and Environmental Research of the US Department of Energy, and in Japan, the Japanese Late Effects Group started similar efforts around the same time to archive the American and Japanese data in the US National Radiobiology Archives (NRA) and the Japanese Radiobiological Archives (JRA), respectively. The result was an aggregated database of primary data from European, Japanese and US sources, the International Radiation Archive (IRA) (Gerber et al. 1999). The JANUS data and Argonne beagle data held at Northwestern University (NURA archive) were also included. The resulting collection of datasets contains nearly all radiation biology studies using animals carried out between 1960 and 1998 in Europe, the US, and Japan, involving a total of more than 400,000 animals (Gerber et al. 1996, Gerber and Wick 2004) (see Table 2). This exercise in international data acquisition and curation was begun by Dr. George Gerber but was picked up in a formal project funded by the European Commission in 2006 when it was decided to integrate all of the data across datasets (Gerber et al. 2006). By then the data had been included in a simple non-relational database and had been hand curated from the original sources. In some cases these were institutional reports, but in others punched card and IBM tapes were transcribed. This raised multiple problems. Firstly, that of the accuracy of transcription was uncertain. More importantly, the lack of standardisation, particularly in animal histopathological diagnoses created a problem.

A variety or terminologies were used in the contributing datasets. SNOMED, ICD9 for humans (used for the data of the few human studies included), a local derivation of SNOMED "SNODOG" for beagles, and local institutional nomenclatures particularly for mice (DIS-ROD). In order to harmonise these classifications a pathology committee was established containing histopathologists from Europe, Japan and the USA in order to assess the correspondence of terms for the same lesion and in some cases to review slides where they were available to confirm the correspondence between the legacy term and a modern term. At that time the MPATH ontology for

mammalian/mouse histopathology had recently been developed as part of the Pathbase database (Schofield et al. 2004) (Schofield et al. 2004, Schofield et al. 2013) and it was decided that this and the combination of anatomical terms available from the mouse anatomy ontology (Hayamizu et al. 2005) should be the basis for standardisation (Tapio et al. 2008). Human disease terms were translated into current ICD 10 classes. The advantage of using the MPATH ontology was beginning to become apparent at the time the curation exercise was undertaken as it was beginning to be widely adopted elsewhere and allowed not only for integration and aggregation of datasets within ERA but also for programmatic access from the outside of the database and in principle integration with external datasets (Birschwilks et al. 2011). The ability for query extension and subsumption over the ontology proved useful, but to date the full analysis ability provided by the ontology coding has not been exploited. A technical feature of interest is that the data curated in the NURA from the JANUS experiments was integrated programmatically into ERA, a task that was feasible only because the NURA archives were based on a modern relational database platform. The integrity and accuracy of data entry to the database was sampled and hand-checked, with precise estimates of error rate and expert evaluation.

The experience gained in creating the ERA database is applicable to any manually curated aggregation and integration of legacy data. All data are available from the ERA website held by the Federal Office for Radiation Protection (Federal Office for Radiation Protection) (http://www.bfs.de/EN/bfs/science-research/projects/era/era_node.html). Already the data have been used to validate DDRF estimates (Haley et al. 2015) and to describe relevant doses and dose-rates (Ruhm et al. 2018) in radiation protection; further studies on the aggregated data are planned.

Several institutions still have ongoing large-scale programmes of rodent exposure studies and have created institutional databases for their primary data collected over several decades. Notable amongst these are two Japanese institutions: The Institute of Environmental Studies and the Japanese Institutes for Quantum and Radiological Science and Technology (QST) with the National Institute of Radiological Sciences being now part of it.

Institute of Environmental Studies database

For the past 22 years the Institute of Environmental studies (IES) in Rokkasho, Aomori prefecture, Japan has been studying the biological effects of long-term external exposure in mice (Braga-Tanaka et al. 2018). The facility at the IES is important in its ability to deliver low doses over the complete lifetime of an experimental animal. Dose rates of 0.05-1 mGy over 400 days, comparable to the doses accumulated by radiation workers. Mice subsequently analysed for lifespan compromise, cancer and other disease plus chromosome abnormalities and transgenerational effects. The accumulated datasets and biological specimens represent a major resource for chronic dose effect assessment and have an important input into the determination of safety limits and risk, especially for occupational exposure.

QST-NIRS J-SHARE database

The Japanese Institutes for Quantum and Radiological Science and Technology (QST) and its National Institute of Radiological Sciences have had a program of large-scale external exposure of rodents to a variety of radiation qualities, X ray, γ, neutron and heavy ions using the Heavy Ion Medical Accelerator at Chiba (HIMAC). These experiments focus on cancer research and many of the experiments are lifespan studies, mostly on wild type or genetically manipulated inbred mice and rats. The QST-NIRS has decided to make their accumulated primary data on more than 13, 000 animals available through the J-Share database (Morioka et al., 2019). This database adds to the very large-scale mouse experiments discussed above with the exception not only containing legacy data but in principle will be augmented by new data from ongoing experiments.

German rodent Thorotrast experiments

Four different studies were conducted at the Deutsches Krebsforschungszentrum, Heidelberg, Germany, from the years 1975 – 1989 examining the effects of exposure of rats to Thorotrast agent (Wegener et al., 1983). The main aims were to determine carcinogenicity and the respective roles of the radioactive and chemical component in

Thorotrast gel-induced tumours. The administration of Thorotrast led to lifelong chronic alpha-particle irradiation by thorium decay products, mainly in the organs of deposition. A database of the results from these studies can be found in STORE (DOI:10.20348/STOREDB/1133/1199).

Environmental and ecological data

The need to develop and sustain competence and experimental infrastructures for radioecology in Europe has become an increasingly urgent need. First addressed through the creation of the European Radioecology Alliance in September 2012, which was officially formed as an association in September 2012, the ALLIANCE now consists of 27 members from 14 European countries. Under the auspices of ALLIANCE the Network of Excellence (NoE), Strategic Network for Integrating Radioecology (STAR), was funded in 2011 by the European Commission as part of Framework Programme 7 (FP7). The framework and strategic plan developed under STAR continued under COMET **(**COordination and iMplementation of a pan-European instrumenT for radioecology**)** in 2013, a combined Collaborative Project and Coordination and Support Action under the EC/ Euratom FP7. As part of the integrative and educational infrastructure developed under COMET the radioecology exchange was created to act as a repository and portal for radioecological data (Muikku et al. 2018). This centralization, sharing and dissemination of large datasets is an established norm within the ecological community, and the coordination shown under ALLIANCE is a model for other communities to follow. The UK Natural Environment Research Council (NERC) developed a similar data centre in the early 2000s which has now become the NERC Environmental Information Data Centre (NERC Data_Centre), itself currently containing 15 radioecology datasets (data accessed 5.11.18). The Radioecology Exchange contains a wide range of datasets from six European countries and Japan from the STAR NoE and is a key resource in radioecology (https://radioecologyexchange.org/content/radioecology-data).

Of other radioecology databases of note, the FREDERICA database (Copplestone et al. 2008) contains data on the effects of radiation on non-human biota curated from the scientific literature. The data contains, amongst other elements, details of

exposures, biological effects, environmental conditions, life cycle, pathway of exposure etc. It currently contains approximately 30,000 expert-curated data entries from around 1200 papers. The wildlife transfer database (Copplestone et al. 2013) (http://www.wildlifetransferdatabase.org/) provides parameter values for use in environmental radiological assessments to estimate the transfer of radioactivity to non-human biota. The PROBA UIAR database contains radionuclide spatial distribution data from the Chernobyl exclusion zone (Kashparov et al. 2018) and can be found both in the NERC datacentre (Kashparov et al., 2017) and the STORE database (http://dx.doi.org/DOI: 10.20348/STOREDB/1087).

Biological and inorganic sample archives

Most of the repositories of materials, both biological and non-biological are associated with large-scale data collection exercises. In many cases the materials were collected to permit measurement of levels of contamination, but also for histopathological and molecular investigation. In some cases, material can be utilised for purposes that were not foreseen at their collection, particularly molecular analyses (Tapio and Atkinson 2008) and there are examples of these from the Mayak tissue bank (Azimzadeh, et al. 2017), the NURA (Haley et al. 2011, Paunesku et al. 2012) and the Chernobyl Thyroid Tissue bank (Abend et al. 2013).

The Chernobyl Tissue Bank, UK

The Chernobyl Tissue Bank (CTB) is an international cooperation which was established in 1998 and which is coordinated by Imperial College London, UK (Thomas 2012). It collects, stores and distributes biological samples from patients with thyroid carcinomas and cellular adenomas who were exposed as children or juveniles by fallout from the Chernobyl accident and resident in contaminated regions of Ukraine and Russia. In addition to the biobanks the CTB keeps information on the patients. It also houses research data derived by researchers using the CTB biomaterials. Data and biomaterial can be accessed once the request is approved in a standard application process.

WISMUT archive, Germany

Along with data from the German uranium miners cohort, a bank with biological samples from former uranium miners and healthy controls was established as a part of an international project (Rosenberger et al. 2018). Information on the German miners is kept in STORE and can be accessed on request (http://dx.doi.org/doi:10.20348/STOREDB/1034).

Radiation Effects Research Foundation (RERF), Japan

A sub-cohort of 15, 000 individuals of the LSS of atomic bomb survivors, the Adult Health Survey, has used biennial health surveys to follow up on all morbidities in addition to cancer. Biological samples have been collected including serum, plasma, urine, lymphocytes, paraffin embedded tissue blocks, prepared slides, and teeth. In 2013, the Biosample Center (RP3-15) was established at RERF with the aim of archiving and curating these biological samples. One of the aims of the project is to consider how to make the samples available to the wider community through collaborations. This involves many complex ethical, legal, and political considerations but it is clear that this is an invaluable resource, which will soon be exploited to improve our understanding of radiation-associated disease mechanisms using new technologies through collaborative studies.

The RERF coordinated study of TEPCO emergency workers from Fukushima discussed above (Kitamura, Okubo and Kodama 2018), is collecting blood and urine from subjects from each of the local medical institutions. Frozen biomaterial exists and plans how to use the samples or make them available in the future are under development.

Southern Urals Biophysical Institute (SUBI), Russia

The research at SUBI, conducted between 1949 to 1996, included studies of alpha- (^{234,235}U, ^{237}Np, $^{238, 239}$Pu, ^{241}Am) and beta- (^{3}H, ^{90}Sr, ^{137}Cs, ^{144}Ce) emitters delivered via different routes into a range of species including rodents (mice, rats) and rabbits,

and other mammals (dog, pig, monkey). Biological material was obtained from more than 23,000 animals; much of it preserved in the SUBI Radiobiological Archive (Abbott 2012). A large amount of the biomaterial is still uncurated and difficult to attribute to the individual animal, but at least for six selected experiments with rodents (mostly Wistar rats) the biomaterial was catalogued and the experiments were described in detail. Information about the experiments with more than 6,000 animals, corresponding to the amount of available samples, and ways of how to get access to these are described in STORE (http://dx.doi.org/doi:10.20348/STOREDB/1056).

Human material is also archived in SUBI. The Russian Radiobiological Human Tissue Repository (RHTR) was established to collect and store biological samples relevant to the human health effects of chronic, low-dose radiation exposure (Loffredo et al. 2017). The RHTR enrolled two cohorts between 1951 and 2017: exposed workers at the Mayak facilities and, as controls, local residents who were never occupationally exposed to ionizing radiation. These samples are annotated with demographic, occupational, dosimetric and medical information. The repository consists of surgical tissues from 900 individuals, autopsy samples from an additional 1000, together with blood samples and DNA from family trios. Both specimens and data are available to the community.

Radiobiological Archive of Large-scale Animal Experiments at QST-NIRS: J-SHARE, Japan

The J-SHARE project described above also includes an extensive archive of biological specimens. To date these consist of material from:

- Lifespan studies of 10,220 B6C3F1 male and female mice at different life stages, irradiated with gamma rays, carbon ions and neutrons.
- Studies on mammary gland and lung carcinogenesis with 2,200 Sprague Dawley female rats and 1,429 Wistar female rats, respectively.
- Studies on brain, digestive tract and renal tumorigenesis utilizing genetically-modified animals.

- Studies on the combined effect of radiation and chemicals.
- Studies for anticarcinogenic properties of caloric restriction and specific antioxidant nutrients and phytochemicals.

Frozen samples are retained along with experimental protocols, paraffin blocks, and histopathological slides. Digitisation of slides is being carried out to produce an archive of zoomable images using the Hamamatsu NanoZoomer. Embedded and frozen tissues are available for molecular analysis.

Institute of Environmental sciences – IES, Rokkasho, Japan

The IES has been conducting studies especially on low-dose chronic irradiation for the last 21 years. Much of the material from these experiments has been archived, mainly as formalin fixed paraffin embedded materials but also frozen (Braga-Tanaka, et al., 2018). This constitutes a major resource of well-preserved and characterised materials from low dose irradiation experiments.

Sample bank of Fukushima animals, Japan

Following the Fukushima Daiichi Nuclear Power Plant accident, a sample bank of animals affected was established. Domestic livestock were collected from the evacuation zone of August 29, 2011 and organs were sampled, and either stored as formalin fixed, paraffin embedded blocks or frozen at -80C (Takahashi et al. 2015). As of the end of March 2015, organs (1,270) and peripheral blood samples (200) from 302 exposed cows had been archived, and analysis on radionuclide content carried out (Fukuda et al. 2013). More recently the sample bank has been augmented by the collection of organs from more than 400 Japanese macaques (Urushihara et al. 2018 and M. Fukumoto. Pers. Comm.). Detailed environmental dosimetry, geographical distribution and other data are available on request.

The National Human Radiobiology Tissue Repository, USTUR, USA

The National Human Radiobiology Tissue Repository (NHRTR) within the United

States Transuranium and Uranium Registries (USTUR) holds around 9,000 frozen and formalin-fixed tissue samples from 40 whole- and 92 partial-body USTUR donors, and around 10,000 acid-digested tissue samples for radioactivity determination (Tolmachev et al. 2011). The role of USTUR, a US federally funded institution, is to study the biokinetics and internal dosimetry of actinides in occupationally exposed individuals who volunteer their post-mortem tissues for scientific use. NHRTR also houses historical frozen, ashed, dried, and plastic-embedded bone samples from the radium studies carried out by Argonne National Laboratory, the Massachusetts Institute of Technology, and the New Jersey Radium Research Project. It also houses the materials from the historic Radium dial painters studies (see https://ustur.wsu.edu/nhrtr/). Materials are freely available subject to ethical and legal permissions.

The Nagasaki Atomic Bomb Survivors' Tumor Tissue Bank, Japan

Beginning in April, 2008, a cohort study has been initiated at Nagasaki University — the Global Strategic Center for Radiation Health Risk Control—to analyse solid cancers and haemopoietic malignancies, radiation exposure information, and clinical data collected from atomic bomb survivors in Nagasaki (Miura et al. 2015). Tumour and surrounding normal tissue are removed at surgery and archived together with personal, historical dose and demographic data. Between 2008 and 2015 around 600 samples were archived, and DNA and RNA prepared.

Northwestern University Radiation Archives (NURA), USA

As described above much of the data from the Argonne experiments on beagles and the JANUS rodent irradiation studies has now been archived and curated at the Northwestern University Radiation Archive (NURA). Along with this data paraffin-embedded material is archived both for the beagle experiments (janus.northwestern.edu/dog_tissues) and Janus mouse experiments (selected tissues; lung, liver, spleen, kidney, heart and gross lesions) along with detailed primary histopathological pathological data from 19, 000 animals (Haley, et al. 2011, Wang, Paunesku and Woloschak 2010). Many of the paraffin embedded tissue samples and original source data are available upon request.

Radioecological and environmental samples

The STAR radioecology project has collated unique data on sample archives throughout Europe, which include samples derived from air (mainly filters), water, soil and building materials, as well as biological material. The data records for these archives may be found on https://radioecology-exchange.org/content/sample-archives along with the appropriate contact details.

Data sharing and archiving platforms

Open data and the sharing imperative

The primary data produced in the course of publicly-funded science represents a common asset for society as much as the analysed and interpreted results. Recent years have seen a unanimous agreement that such data and discoveries should be as accessible as possible by other scientists and the members of society in order to extract the maximum value from that investment. The concept of the science commons is well established and legal economic and social aspects of the commons are the subjects of intensive interest and examination (Cook-Deegan 2007, Mishra and Bubela 2014). We have discussed above the importance of the reuse and sustainability of large individual datasets and aggregates of legacy data. There is currently increasing political and scientific concern about the issues of data sharing, accountability and reproducibility in the biological sciences (Baker 2016, Begley and Ellis 2012, Collins and Tabak 2014), and the preservation and sharing of individual datasets from current studies, especially that data which supports the conclusions of publications. In response, many journals and funding agencies have recently adopted or mandated guidelines for the openness and reuse of primary data and computer code, as well as open access publications (Berg 2018, Federer et al. 2018, Nature editorial 2016, Nature editorial 2017, Stodden et al. 2018). Open data provides better value to society from data reuse, reanalysis, reduction in both duplication and animal experimentation (3Rs) and it improves reproducibility and accountability for claims made in publications.

In response to these developments a framework for data sharing has been established through a consensus process involving investigators, funding agencies, learned societies and journals. The resulting FAIR guidelines for Open Data (Wilkinson et al. 2016) have now been adopted by most major funding agencies, the European Commission and formally by the countries of the OECD and G20 group of nations, to represent a benchmark for open scientific data (Arzberger et al. 2004, Mons et al. 2017). Findability, Accessibility, Interoperability, and Reusability represent the four principles of Open data and are underpinned by, and inseparable from, effective data governance and management and mediated by an open infrastructure (Sansone et al. 2018). The implications of the FAIR guidelines are that data should be discoverable and accessible by a human or by machine, that it should have sufficient metadata to be understandable and implementable and critically that the originator of the data should not be involved in the decision as to whom it is made available. The FAIR principles do not preclude licensing or reasonable charges for access, so FAIR does not necessarily mean free or free of constraint over use, but that data should be accessible under reasonable conditions and in fact most of the data that concerns us lies within the pre-competitive space in any case. Major funding agencies such as the European Commission (Horizon 2020 guidelines, 2016) , and the NIH are now also trialing a FAIR data commons policy (National Institutes of Health, 2018).

Databases and repositories are the essential infrastructure for the research commons and require coordinated development and sustainable funding (Sansone, Cruse and Thorley 2018, Schofield et al. 2010). There already exist large public databases dedicated to particular domains or data types, such as Array Express (RNA expression studies (Kolesnikov et al., 2015)), PRIDE (proteomics (Jarnuczak and Vizcaino, 2017)), and Mouse genome infomatics (MGI), (genomic, variant and phenotypic data on mice (Eppig, 2017)). In Europe many of these core resources have been adopted under the umbrella of the ELIXIR life science informatics infrastructure (Durinx et al. 2016). While the outputs of radiobiology in these areas might be deposited in these databases it became clear that there would be advantages, particularly with regard to the FAIR criteria, for there to be an open, aggregating data platform where any kind of data relating to radiobiology and epidemiology might be archived. These considerations lead to the development of the STORE database.

STORE DB; a database for radiobiology, radioecology and epidemiology

Development of the STORE database began in 2009 under European Commission funding to encourage public data sharing and reuse in the domain of radiation biology. It was sustained through successive grants and was opened to public use in 2014 (see Figure 1A). Now open to public use for three years STORE provides a data type agnostic platform for all kinds of data, ranging from epidemiology and human cohort data to 'omics, cytogenetics, computer code and documents. File structures in STORE are based on the "project" as the top level entity. This forms an envelope for datasets and individual data items in a nested fashion. This means that all of the different types of data associated with a particular project or undertaking can be clustered together to make a coherent set of elements, while each file can be searched and retrieve separately. This clustering of data has distinct advantages over the approaches taken commercial data-agnostic repositories that are centred only on the data entry itself.

Data and datasets are tagged with metadata terms taken from the Ontology for Biomedical Investigations (Bandrowski et al. 2016) and the Experimental factor ontology (Malone et al. 2010), though there are ongoing efforts to augment these ontologies with terms for radiation biology specifically and where there are gaps in term provision these are provided by an in house vocabulary. Current efforts are focussed on creating a radiobiology and epidemiology ontology.

Increasingly STORE is being used by large distributed projects to coordinate and archive primary and derivative data which is then used for support of publications. STORE provides persistent digital object identifiers and accession IDs which use a persistent namespace formally registered with identifiers.org at the EBI. Similarly registered with the FAIRSharing initiative (McQuilton et al. 2016) and *re3data* (Pampel et al. 2013) STORE is a well recognised and accepted data repository. The database is physically located at the BfS in Neuherberg and has the full security of a German Federal data service. The BfS has undertaken to maintain the database indefinitely which means that data will be secure and accessible for the foreseeable

future. Currently STORE contains around 3000 data objects across a wide range of data types; the number is increasing rapidly.

The aim of STORE is to promote open access and reuse of data, as well as the archiving of at-risk or legacy data, thus promoting and enhancing the scientific commons. Consequently, deposition and access to data are free to individual investigators and to funding agencies. Data will be stored live for a guaranteed period of 7 years after the most recent access, after which it will be stored successively for another 7 years and so forth. If data is not accessed for longer than this period then it will be taken offline and stored in "cold storage" or archived to permanent and less expensive media (Schatz 2015). STORE is available on http://www.storedb.org and access is provided by users' ORCID IDs through an intuitive web interface (Figure 1B), although programmatic access is also planned in the near future, compliant with aspirational goals for FAIRing data (Wilkinson et al. 2018).

Current challenges for Open data

Despite widely publicised concern about the availability of data, the adoption of Open data guidelines by funding agencies and increasingly insistence by journals that data supporting the claims made in publications, together with resources such as antibodies and mice, be made publicly available at the time of publication (for details of policies see: (McQuilton, et al. 2016)) there remain significant problems. Analyses indicate that there is still a profound resistance amongst the biomedical community to sharing primary data, even if recommended or mandated by funder or journal. In support of the aims of transparency and reproducibility many bodies have adopted the FAIR guidelines and journals are increasingly modifying their policies to conform to the criteria laid out in the TOP (Transparency, Openness and Reproducibility) guidelines (Nosek et al. 2015). So far, however, evidence suggests that the impact on the culture of data sharing has been slight, with the exception of some journals, such as the PLoS stable (Bloom et al. 2014) where these seems to have been a small but significant impact on the availability of data behind publications (Federer, et al. 2018) in comparison with 2009 when an analysis of PLoS journals came to the conclusion that "our findings suggest that explicit journal policies requiring data sharing do not lead to authors making their data sets available to independent investigators" (Savage and

Vickers 2009).

In 2016 a study looking at the availability of transparent protocols and data in 441 journal articles found that not a single paper made all the raw data available, in contravention of stated journal policies in many cases, and only one made protocols available (Iqbal et al. 2016). In a retrospective study on the 111 most influential articles in psychology and psychiatry, data could only be retrieved in 34% of cases and of these it was often incomplete or otherwise carried restrictions on use or analysis (Hardwicke and Ioannidis 2018). The same authors also examined clinical trial data from PLoS Medicine and disappointingly found only 46% of papers making data available, in journals with apparently stringent data sharing policies (Naudet et al. 2018). A similar proportion of data from ecological studies – 56% – was also found to be incomplete and much unusable (Roche et al. 2015). Attempts are being made to produce guidelines for clinical trial data reporting but as yet there remain difficulties in making these mandatory (Taichman, et al. 2017). While similar surveys have not yet been completed in the domain of radiobiology and epidemiology, it is disappointing that of 14 journals that take significant numbers of papers in radiation biology only one had any stipulation about data availability; Radiation and Experimental Biophysics.

Sharing of human clinical and personal data

Investigators frequently consider data generated in the conduct of a clinical trial or epidemiological study to be effectively proprietary, either belonging to the funder, whether a public or private agency, or to the researchers. The consequence of this is that much valuable data has not been made available for further studies and its full value not realised. In addition to the inability to replicate analyses, this undermines both trust and accountability. The problems of sharing personal data such as genomic sequences or clinical data are complex, but dependent on the exact form of consenting carried out for the study. It is possible to share anonymised personal level data widely, so long as consenting is done appropriately and data held and transferred in a robust encrypted format; exemptions exist in European and other data protection law for the sharing of anonymised health data where that sharing is in the public interest (discussed in (Rumbold and Pierscionek 2017). Radiation epidemiology data is no

exception to this general problem and much of the epidemiological data discussed above is not readily available to researchers. However, genomic and phenotypic data are now widely shared around the world; for example, the UK Biobank project has successfully shared personal data and genomes for more than 100, 000 individuals globally, and the CINECA consortium has launched an infrastructure for the sharing of 1.4 million personal genomes[3]. Some radiation epidemiology cohorts, such as those for the WISMUT miners, were consented with some foresight, and are available on request. However much legacy data cannot be retrospectively re-consented and in those cases access and reuse will inevitably be limited.

The key problem seems not to be current legal constraints on data sharing, but the wide range of approaches and procedures adopted locally by clinical trial and epidemiological units, e.g. (Hopkins et al. 2016). There is a clear need for homogenisation and policy recommendations to ensure adherence to consistent best practice to ensure maximisation of data sharing and exploitation.

An infrastructure for data sharing and archiving

The development of supplementary information sites for journals over the last 20 years is no longer regarded as an adequate repository for primary data, as many of these repositories are unstructured, unstable – data is often lost (Alsheikh-Ali et al. 2011, Anderson, et al. 2006), undiscoverable or not actually submitted, in contradiction to explicit journal policies (Federer, et al. 2018). Moreover, in many cases there is insufficient information attached to data files to allow them to be used for reanalysis or reuse. Where studies have been done on data retained by authors it seems clear that there is a high risk of data "disappearing" (Savage and Vickers 2009) and a recent retraction from Science (Roche 2017) underlines the importance of formal structured and sustainable repositories. It is clear therefore that stable repositories, such as provided by STORE and other public databases form an essential part of the data infrastructure in the biomedical sciences.

[3] https://www.ebi.ac.uk/about/news/press-releases/CINECA-facilitates-transcontinental-human-data-exchange

Why the failure to share?

The impact of cryptic data – ie that which is not available for scrutiny – certainly contributes to lack of reproducibility in the life sciences, the consequences of which are huge cost both to the public purse and to industry, together with delays in delivering the products of the scientific endeavour to the public (Macleod et al. 2014). This in turn has knock-on effects on the political and societal confidence in the scientific enterprise (Piwowar 2011). This is particularly an issue within the biological radiation sciences and radiation protection, where public safety rests so much on the reliable results of research. Availability of data and materials collected as part of a study can have huge added value if reused and subjected to reanalysis as is shown in many examples discussed above in this review. We must question why the sharing of data particularly is so poor.

The UK Joint Information Systems Committee (JISC) has recently conducted a comprehensive survey into the implementation of FAIR principles in the biomedical sciences which comes to very similar conclusions to previous surveys of attitudes in specific disciplines (Allen and Hartland 2018, Blumenthal et al. 2006, Piwowar 2011, Tenopir et al. 2011, Tenopir et al. 2015). One worrying observation is that data from marginally significant or poorly reported experiments seems to dominate the data sharing deficiency (Wicherts et al. 2011), suggesting that there is concern amongst some authors that their data are not checked or reanalysed. There are also the issues of fear of being scooped or of giving help to the competition, and perceived, but often not real, fears about losing the opportunity to protect intellectual property. Similar issues in the radiation biology community are shown by a recent study carried out within the MELODI low dose radiation protection programme (Madas and Schofield 2019). Issues about training in data management, the cost of preparing and submitting data are found in all the studies reported, but an overarching problem, that of data ownership and the personal interests of the investigator are a persistent theme. As crisply summarised by Richard Smith, former editor of the Lancet:

> "Most scientific studies are wrong, and they are wrong because scientists are interested in funding and careers rather than truth" (Smith 2013)

The impact of career incentives on the quality of science is discussed recently in (Smaldino and McElreath 2016). It is clear that training of young investigators, and normalisation of the expectation of open data and transparency should be goals as significant as funding body or journal policy development, and as critical as the stable provision of infrastructure for the preservation and dissemination of publicly funded data.

Challenges for the future

It is not possible in a survey of the data landscape of radiation biology over the past 60 years and more to miss how important has been the critical importance of freely accessed and sustainable archived data. As attitudes change and data floods into the scientific community, we face not only sustainability challenges but challenges in training; both in the data management skills expected of investigators, and the ethics of scientific investigation. Within the scope of the current commentary we cannot claim to have included all of the datasets currently available in radiation biology and epidemiology. We welcome further suggestions from readers, and submissions to the STORE database.

The first challenge of the next 60 years will be how to manage, exploit and, increasingly how to find data. The latter is an informatics challenge already being addressed in the FAIR framework from a technical point of view, but familiarity with informatics as part of normal scientific training is going to become much more important in the imminent future than it ever has been before.

Sustaining the infrastructure for data and biomaterial archiving is the second major challenge. There are several models for the financial and scientific sustainability of databases (Chandras et al. 2009, Kaiser 2016, Reiser et al. 2016, Sansone, Cruse and Thorley 2018, Schofield, et al. 2010), of which none are “one size-fits-all”, and it remains to be seen how the international community grasps this particular nettle with the aim of producing the stable and long term investment in infrastructure that the world scientific community requires. Data, like radiation, does not respect

international boundaries. Without such investment long term the rich data accumulated and accumulating in radiation biology are at risk.

Acknowledgements

Thanks go to the partners of the initial STORE project (HMGU, Germany; SUBI, Russian Federation; The University of Edinburgh-Edinburgh Cancer Research Centre, UK; Imperial College London, UK; RIVM, The Netherlands; IBBL, Luxemburg). Further thanks go to Mandy Birschwilks, Simon Bouffler, Janet Tawn, Gayle Woloschak, Shin Saigusa, Nick Beresford, and Jonathan Bard for their advice, skill and wisdom. Special thanks go to Michael Gruenberger for the programming of STORE, and to Balazs Madas for initiating the CONCERT survey on attitudes to open data. The authors are deeply indebted to Bundesamt fuer Strahlenschutz for guaranteeing the sustainability of STORE.

Disclosure Statement

The authors report no conflicts of interest

Funding

The ERA database was developed within the Euratom FP6 contract 28725 (ERA-PRO). The development of the STORE database was funded by Euratom FP7 contract 232628 (STORE), and partly by 249689 (DoReMi). Current funding is from the Euratom research and training programme 2014-2018 under grant agreement No 662287 (CONCERT - European Joint Programme for the Integration of Radiation Protection Research).

References

1. Abbott A. 2012. Radiation risks: Raiders of the lost archive. Nature.485:162-163.

2. Abend M, Pfeiffer RM, Ruf C, Hatch M, Bogdanova TI, Tronko MD, Hartmann J, Meineke V, Mabuchi K, Brenner AV. 2013. Iodine-131 dose-dependent gene expression: alterations in both normal and tumour thyroid tissues of post-Chernobyl thyroid cancers. British journal of cancer.109:2286-2294. Epub 2013/09/17.

3. Akleyev AV, Krestinina LY, Degteva MO, Tolstykh EI. 2017. Consequences of the radiation accident at the Mayak production association in 1957 (the 'Kyshtym Accident'). J Radiol Prot.37:R19-R42. Epub 2017/07/14.

4. Allen R, Hartland D. 2018. FAIR in practice - Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles. Available from: https://zenodo.org/record/1245568 - .W0c-ztX0mUl

5. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. 2011. Public availability of published research data in high-impact journals. PloS one.6:e24357.

6. Anderson N, Tarczy-Hornoch P, Bumgarner R. 2006. On the persistence of supplementary resources in biomedical publications. BMC bioinformatics.7:260.

7. Apsalikov K, Lipikhina A, Grosche B, Belikhina T, Ostroumova EV, Shinkarev S, Stepanenko V, Muldagaliev T, Yoshinaga S, Zhunussova T, et al. 2019. The State Scientific Automated Medical Registry, Kazakhstan: An important resource for low-dose radiation health research. Radiation and environmental biophysics. (In Press).

8. Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P. 2004. An International Framework to Promote Access to Data. Science.303:1777.

9. Azimzadeh O, Azizova T, Merl-Pham J, Subramanian V, Bakshi MV, Moseeva M, Zubkova O, Hauck SM, Anastasov N, Atkinson MJ, et al. 2017. A dose-dependent perturbation in cardiac energy metabolism is linked to radiation-induced ischemic heart disease in Mayak nuclear workers. Oncotarget.8:9067-9078. Epub 2016/07/09.

10. Azizova TV, Day RD, Wald N, Muirhead CR, O'Hagan JA, Sumina MV, Belyaeva ZD, Druzhinina MB, Teplyakov, II, Semenikhina NG, et al. 2008. The "clinic" medical-dosimetric database of Mayak production association workers: structure, characteristics and prospects of utilization. Health physics.94:449-458. Epub 2008/04/12.

11. Baker M. 2016. 1,500 scientists lift the lid on reproducibility. Nature.533:452-454.

12. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, Clancy K, Courtot M, Derom D, Dumontier M, et al. 2016. The Ontology for Biomedical Investigations. PloS one.11:e0154556.

13. Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. Nature.483:531-533.

14. Berg J. 2018. Progress on reproducibility. Science.359:9.

15. Bernier MO, Baysson H, Pearce MS, Moissonnier M, Cardis E, Hauptmann M, Struelens L, Dabin J, Johansen C, Journy N, et al. 2018. Cohort Profile: the EPI-CT study: A European pooled epidemiological study to quantify the risk of radiation-induced cancer from paediatric CT. Int J Epidemiol. Epub 2018/11/06.

16. Birschwilks M, Gruenberger M, Adelmann C, Tapio S, Gerber G, Schofield PN, Grosche B. 2011. The European radiobiological archives: online access to data from radiobiological experiments. Radiation research.175:526-531. Epub 2011/01/27.

17. Bloom T, Ganley E, Winker M. 2014. Data Access for the Open Access Literature: PLOS's Data Policy. PLoS Biol.12:e1001797.

18. Blumenthal D, Campbell EG, Gokhale M, Yucel R, Clarridge B, Hilgartner S, Holtzman NA. 2006. Data withholding in genetics and the other life sciences: prevalences and predictors. Acad Med.81:137-145.

19. Boice JD, Jr., Ellis ED, Golden AP, Girardi DJ, Cohen SS, Chen H, Mumma MT, Shore RE, Leggett RW. 2018. The Past Informs the Future: An Overview of the Million Worker Study and the Mallinckrodt Chemical Works Cohort. Health physics.114:381-385. Epub 2018/02/27.

20. Braga-Tanaka I, Tanaka S, Kohda A, Takai D, Nakamura S, Ono T, Tanaka K, Komura J. 2018. Experimental studies on the biological effects of chronic low dose-rate radiation exposure in mice: overview of the studies at the Institute for Environmental Sciences. International journal of radiation biology.94:423-433.

21. Cardis E, Hatch M. 2011. The Chernobyl accident--an epidemiological perspective. Clin Oncol (R Coll Radiol).23:251-260. Epub 2011/03/15.

22. Chandras C, Weaver T, Zouberakis M, Smedley D, Schughart K, Rosenthal N, Hancock JM, Kollias G, Schofield PN, Aidinis V. 2009. Models for financial sustainability of biological databases and resources. Database : the journal of biological databases and curation.2009:bap017.

23. Collins FS, Tabak LA. 2014. Policy: NIH plans to enhance reproducibility. Nature.505:612-613.

24. Cook-Deegan R. 2007. The science commons in health research: structure, function, and value. The Journal of technology transfer.32:133-156. Epub 2006/12/07.

25. Copplestone D, Hingston J, Real A. 2008. The development and purpose of the FREDERICA radiation effects database. Journal of environmental radioactivity.99:1456-1463.

26. Copplestone D, Beresford NA, Brown JE, Yankovich T. 2013. An international database of radionuclide concentration ratios for wildlife: development and uses. Journal of environmental radioactivity.126:288-298. Epub 2013/07/03.

27. Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, Cook C, Gasteiger E, Kim JH, Lopez R, et al. 2016. Identifying ELIXIR Core Data Resources. F1000Research.5. Epub 2017/03/30.

28. Faden R editor 1996. The Human Radiation Experiments; Final Report of the Advisory Committee on Human Radiation Experiments . OUP, New York.

29. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, Thompson H. 2018. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PloS one.13:e0194768.

30. Fry SA. 1998. Studies of U.S. Radium Dial Workers: An Epidemiological Classic. Radiation research.150:S21-S29.

31. Fukumoto M. 2014. Radiation pathology: from thorotrast to the future beyond radioresistance. Pathol Int.64:251-262. Epub 2014/06/27.

32. Fukuda T, Kino Y, Abe Y, Yamashiro H, Kuwahara Y, Nihei H, Sano Y, Irisawa A, Shimura T, Fukumoto M, et al. 2013. Distribution of artificial radionuclides in abandoned cattle in the evacuation zone of the Fukushima Daiichi nuclear power plant. PloS one.8:e54312. Epub 2013/02/02.

33. Gerber G, Watson C, Sugahara T, Okada S. 1996. International Radiobiology Archives of Long-Term Animal Studies I. Descriptions of Participating Institutions and Studies. DOI:10.20348/STOREDB/21119/21185.

34. Gerber GB, Wick RR. 2004. The European Radiobiology Archives (ERA), present state and future developments. Radiation protection dosimetry.112:529-530. Epub 2004/12/30.

35. Gerber GB, Wick RR, Kellerer AM, Hopewell JW, Di Majo V, Dudoignon N, Gossner W, Stather J. 2006. The European Radiobiology Archives (ERA)--content, structure and use illustrated by an example. Radiation protection dosimetry.118:70-77. Epub 2005/10/26.

36. Gerber GB, Wick RR, Watson CR, Gossner W, Kellerer AM. 1999. International radiobiology archives of long-term animal studies: structure, possible uses and potential extension. Radiation and environmental biophysics.38:75-79. Epub 1999/08/26.

37. Grosche B, Birschwilks M, Wesch H, Kaul A, van Kaick G. 2016. The German Thorotrast Cohort Study: a review and how to get access to the data. Radiation and environmental biophysics.55:281-289. Epub 2016/05/08.

38. Grosche B, Lackland DT, Land CE, Simon SL, Apsalikov KN, Pivina LM, Bauer S, Gusev BI. 2011. Mortality from cardiovascular diseases in the Semipalatinsk

historical cohort, 1960-1999, and its relationship to radiation exposure. Radiation research.176:660-669. Epub 2011/07/27.

39. Haley B, Wang Q, Wanzer B, Vogt S, Finney L, Yang PL, Paunesku T, Woloschak G. 2011. Past and future work on radiobiology mega-studies: a case study at Argonne National Laboratory. Health physics.100:613-621.

40. Haley BM, Paunesku T, Grdina DJ, Woloschak GE. 2015. The Increase in Animal Mortality Risk following Exposure to Sparsely Ionizing Radiation Is Not Linear Quadratic with Dose. PloS one.10:e0140989. Epub 2015/12/10.

41. Hamra GB, Richardson DB, Cardis E, Daniels RD, Gillies M, O'Hagan JA, Haylock R, Laurier D, Leuraud K, Moissonnier M, et al. 2016. Cohort Profile: The International Nuclear Workers Study (INWORKS). International journal of epidemiology.45:693-699.

42. Hardwicke TE, Ioannidis JPA. 2018. Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. PloS one.13:e0201856.

43. Hasegawa A, Tanigawa K, Ohtsuru A, Yabe H, Maeda M, Shigemura J, Ohira T, Tominaga T, Akashi M, Hirohashi N, et al. 2015. Health effects of radiation and other health problems in the aftermath of nuclear accidents, with an emphasis on Fukushima. Lancet.386:479-488. Epub 2015/08/08.

44. Hatch M, Cardis E. 2017. Somatic health effects of Chernobyl: 30 years on. European journal of epidemiology.32:1047-1054. Epub 2017/09/21.

45. Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. 2005. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome biology.6:R29.

46. Haylock RGE, Gillies M, Hunter N, Zhang W, Phillipson M. 2018. Cancer mortality and incidence following external occupational radiation exposure: an update of the 3rd analysis of the UK national registry for radiation workers. British journal of cancer. Epub 2018/08/16.

47. Hewlett RG, Anderson OE, Duncan F. 1990. A History of the United States Atomic Energy Commission: The new world, 1939: University of California Press.

48. H2020 Guidelines on FAIR Data Management. 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

49. Hopkins C, Sydes M, Murray G, Woolfall K, Clarke M, Williamson P, Tudur Smith C. 2016. UK publicly funded Clinical Trials Units supported a controlled access approach to share individual participant data but highlighted concerns. Journal of clinical epidemiology.70:17-25.

50. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. 2016. Reproducible Research Practices and Transparency across the Biomedical Literature. PLOS Biology.14:e1002333.

51. Ishikawa T, Yasumura S, Ozasa K, Kobashi G, Yasuda H, Miyazaki M, Akahane K, Yonai S, Ohtsuru A, Sakai A, et al. 2015. The Fukushima Health Management Survey: estimation of external doses to residents in Fukushima Prefecture. Scientific reports.5:12712. Epub 2015/08/05.

52. Kaiser J. 2016. Funding for key data resources in jeopardy. Science.351:14.

53. Kashparov V, Levchuk S, Zhurba M, Protsak V, Khomutinin Y, Beresford NA, Chaplow JS. 2018. Spatial datasets of radionuclide contamination in the Ukrainian Chernobyl Exclusion Zone. Earth Syst Sci Data.10:339-353.

54. Kashparov V, Levchuk S, Zhurba M, Protsak V, Khomutinin Y, Beresford NA, Chaplow JS. 2017. Spatial datasets of radionuclide contamination in the Ukrainian Chernobyl Exclusion Zone. In: NERC Environmental Information Data Centre. https://doi.org/10.5285/782ec845-2135-4698-8881-b38823e533bf (accessed 5.11.18)

55. Kitamura H, Okubo T, Kodama K. 2018. Epidemiological study of health effects in Fukushima nuclear workers – study design and progress report. Radiation protection dosimetry. 2018 Aug 23. doi: 10.1093/rpd/ncy136. [Epub ahead of print]

56. Krestinina LY, Preston DL, Ostroumova EV, Degteva MO, Ron E, Vyushkova OV, Startsev NV, Kossenko MM, Akleyev AV. 2005. Protracted radiation exposure and cancer mortality in the Techa River Cohort. Radiation research.164:602-611. Epub 2005/10/22.

57. Kreuzer M, Auvinen A, Cardis E, Hall J, Jourdain JR, Laurier D, Little MP, Peters A, Raj K, Russell NS, et al. 2015. Low-dose ionising radiation and cardiovascular diseases--Strategies for molecular epidemiological studies in Europe. Mutat Res Rev Mutat Res.764:90-100. Epub 2015/06/05.

58. Kreuzer M, Schnelzer M, Tschense A, Walsh L, Grosche B. 2010. Cohort profile: the German uranium miners cohort study (WISMUT cohort), 1946-2003. Int J Epidemiol.39:980-987. Epub 2009/06/06.

59. Land CE, Zhumadilov Z, Gusev BI, Hartshorne MH, Wiest PW, Woodward PW, Crooks LA, Luckyanov NK, Fillmore CM, Carr Z, et al. 2008. Ultrasound-detected thyroid nodule prevalence and radiation dose from fallout. Radiation research.169:373-383. Epub 2008/03/28.

60. Liu W, Haley B, Kwasny MJ, Li JJ, Grdina DJ, Paunesku T, Woloschak GE. 2013. Comparing radiation toxicities across species: an examination of radiation effects in Mus musculus and Peromyscus leucopus. International journal of radiation biology.89:391-400. Epub 2013/02/01.

61. Loffredo C, Goerlitz D, Sokolova S, Leondaridis L, Zakharova M, Revina V, Kirillova E. 2017. The Russian Human Radiobiological Tissue Repository: A Unique Resource for Studies of Plutonium-Exposed Workers. Radiation protection dosimetry.173:10-15. Epub 2016/11/26.

62. Macklis RM. 1990. Radithor and the era of mild radium therapy. JAMA : the journal of the American Medical Association.264:614-618. Epub 1990/08/01.

63. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, Al-Shahi Salman R, Chan AW, Glasziou P. 2014. Biomedical research: increasing value, reducing waste. Lancet.383:101-104.

64. Madas B, Schofield P. 2019. Survey on data management in radiation protection research. Radiation protection dosimetry. (In Press, https://dx.doi.org/10.1093/rpd/ncy250). Also available as preprint from: arXiv:1805.05463 [cs.CY].

65. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. 2010. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics (Oxford, England). 26:1112-1118. Epub 2010/03/03.

66. McCally M, Cassel C, Kimball DG. 1994. U.S. government-sponsored radiation research on humans 1945-1975. Med Glob Surviv.1:4-17. Epub 1994/03/01.

67. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, Sansone S-A. 2016. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database.2016:baw075-baw075.

68. Mishra A, Bubela T. 2014. Legal Agreements and the Governance of Research Commons: Lessons from Materials Sharing in Mouse Genomics. OMICS: A Journal of Integrative Biology.18:254-273.

69. Miura S, Akazawa Y, Kurashige T, Tukasaki K, Kondo H, Yokota K, Mine M, Miyazaki Y, Sekine I, Nakashima M. 2015. The Nagasaki Atomic Bomb Survivors' Tumor Tissue Bank. Lancet.386:1738. Epub 2015/11/08.

70. Mons B, Neylon C, Velterop J. 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inform Services Use.37:49-56. Epub 49.

71. Morioka, T, Blyth, B.J., Imaoka, T., Nishimura, M., Takeshita, H., Shimomura, T., Ohtake, J., Ishida, A., Schofield,P.N., Grosche, B., Kulka, U., Shimada, Y., Kakinuma, S., and Yamada, Y. 2019, Establishing the Japan-StoreHouse of Animal Radiobiology Experiments (J-SHARE), a large-scale necropsy and histopathology archive providing international access to important radiobiology data . Int. J. Rad. Biol. (Submitted)

72. Muikku M, Beresford NA, Garnier-Laplace J, Real A, Sirkka L, Thorne M, Vandenhove H, Willrodt C. 2018. Sustainability and integration of radioecology-position paper. J Radiol Prot.38:152-163. Epub 2017/11/22.

73. Muller HJ. 1927. Artificial Transmutation of the Gene. Science.66:84-87.

74. National Academy of Engineering 1984. Memorial Tributes: Volume 2 Washington, DC: The National Academies Press. 978-0-309-03482-1

75. National Institutes of Health, 2018. https://commonfund.nih.gov/commons/ (Accessed 5.11.18)

76. Nature Editorial. 2016. Reality check on reproducibility. Nature.533:437. Epub 2016/05/27.

77. Nature Editorial. 2017. Announcement: Towards greater reproducibility for life-sciences research in Nature. Nature.546:8. Epub 2017/06/02.

78. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. 2018. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine BMJ. 360:k400

79. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, et al. 2015. Scientific standards. Promoting an open research culture. Science.348:1422-1425.

80. Ozasa K. 2016. Epidemiological research on radiation-induced cancer in atomic bomb survivors. Journal of radiation research.57 Suppl 1:i112-i117. Epub 2016/03/16.

81. Ozasa K, Grant EJ, Kodama K. 2018. Japanese Legacy Cohorts: The Life Span Study Atomic Bomb Survivor Cohort and Survivors' Offspring. Journal of Epidemiology.28:162-169.

82. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, Goebelbecker H-J, Gundlach J, Schirmbacher P, Dierolf U. 2013. Making Research Data Repositories Visible: The re3data.org Registry. PloS one.8:e78080.

83. Paunesku D, Paunesku T, Wahl A, Kataoka Y, Murley J, Grdina DJ, Woloschak GE. 2008. Incidence of tissue toxicities in gamma ray and fission neutron-exposed mice treated with Amifostine. International journal of radiation biology.84:623-634.

84. Paunesku T, Wanzer MB, Kirillova EN, Muksinova KN, Revina VS, Lyubchansky ER, Grosche B, Birschwilks M, Vogt S, Finney L, et al. 2012. X-ray fluorescence microscopy for investigation of archival tissues. Health physics.103:181-186. Epub 2012/09/07.

85. Piwowar HA. 2011. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PloS one.6:e18657.

86. Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E. 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. Database : the journal of biological databases and curation.2016:baw018.

87. Roche DG, Kruuk LE, Lanfear R, Binning SA. 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biol.13:e1002295. Epub 2015/11/12

88. Roche DG. 2017. Evaluating Science; open-data policy. Science.357:654.

89. Rosenberger A, Hung RJ, Christiani DC, Caporaso NE, Liu G, Bojesen SE, Le Marchand L, Haiman CA, Albanes D, Aldrich MC, et al. 2018. Genetic modifiers of radon-induced lung cancer risk: a genome-wide interaction study in former uranium miners. Int Arch Occup Environ Health.91:937-950. Epub 2018/07/05.

90. Ruhm W, Azizova T, Bouffler S, Cullings HM, Grosche B, Little MP, Shore RS, Walsh L, Woloschak GE. 2018. Typical doses and dose rates in studies pertinent to radiation risk inference at low doses and low dose rates. Journal of radiation research.59:ii1-ii10. Epub 2018/02/13.

91. Rumbold JM, Pierscionek BK. 2017. A critique of the regulation of data science in healthcare research in the European Union. BMC medical ethics.18:27. Epub 2017/04/09.

92. Sadetzki S, Chetrit A, Freedman L, Stovall M, Modan B, Novikov I. 2005. Long-term follow-up for brain tumor development after childhood exposure to ionizing radiation for tinea capitis. Radiation research.163:424-432. Epub 2005/04/01.

93. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML. 2019. Deep learning in medical imaging and radiation therapy. Medical Physics.46:e1-e36.

94. Sansone SA, Cruse P, Thorley M. 2018. High-quality science requires high-quality open data infrastructure. Scientific data.5:180027.

95. Savage CJ, Vickers AJ. 2009. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PloS one.4:e7078.

96. Schatz MC. 2015. Biological data sciences in genome research. Genome research. 25:1417-1422.

97. Schofield PN, Bard JB, Booth C, Boniver J, Covelli V, Delvenne P, Ellender M, Engstrom W, Goessner W, Gruenberger M, et al. 2004. Pathbase: a database of mutant mouse pathology. Nucleic acids research.32:D512-515.

98. Schofield PN, Eppig J, Huala E, de Angelis MH, Harvey M, Davidson D, Weaver T, Brown S, Smedley D, Rosenthal N, et al. 2010. Research funding. Sustaining the data and bioresource commons. Science.330:592-593.

99. Schofield PN, Sundberg JP, Sundberg BA, McKerlie C, Gkoutos GV. 2013. The mouse pathology ontology, MPATH; structure and applications. Journal of biomedical semantics.4:18.

100. Sekiya M, Yamasaki M. 2016. Rolf Maximilian Sievert (1896-1966): father of radiation protection. Radiol Phys Technol.9:1-5. Epub 2015/07/30.

101. Smaldino PE, McElreath R. 2016. The natural selection of bad science. R Soc Open Sci.3:160384. Epub 2016/10/06.

102. Smith R. 2013. Time for science to be about truth rather than careers. . Available from: https://blogs.bmj.com/bmj/2013/09/09/richard-smith-time-for-science-to-be-about-truth-rather-than-careers/

103. Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences of the United States of America.115:2584-2589. Epub 2018/03/14.

104. Stone R. 1993. Scientists study 'cold war' fallout. Science.262:1968. Epub 1993/12/24.

105. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, Hong S-T, Haileamlak A, Gollogly L, Godlee F, et al. 2017. Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. PLOS Medicine.14:e1002315.

106. Takahashi S, Inoue K, Suzuki M, Urushihara Y, Kuwahara Y, Hayashi G, Shiga S, Fukumoto M, Kino Y, Sekine T, et al. 2015. A comprehensive dose evaluation project concerning animals affected by the Fukushima Daiichi Nuclear Power Plant accident: its set-up and progress. Journal of radiation research.56 Suppl 1:i36-i41. Epub 2015/12/18.

107. Tapio S, Atkinson MJ. 2008. Molecular information obtained from radiobiological tissue archives: achievements of the past and visions of the future. Radiation and environmental biophysics.47:183-187.

108. Tapio S, Schofield PN, Adelmann C, Atkinson MJ, Bard JL, Bijwaard H, Birschwilks M, Dubus P, Fiette L, Gerber G, et al. 2008. Progress in updating the European Radiobiology Archives. International journal of radiation biology.84:930-936.

109. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M. 2011. Data sharing by scientists: practices and perceptions. PloS one.6:e21101.

110. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PloS one.10:e0134826.

111. Thomas GA. 2012. The Chernobyl Tissue Bank: integrating research on radiation-induced thyroid cancer. J Radiol Prot.32:N77-80. Epub 2012/03/08.

112. Tolmachev SY, Ketterer ME, Hare D, Doble P, James AC. 2011. The US Transuranium and Uranium Registries: forty years' experience and new directions in the analysis of actinides in human tissues. Radiochimica Acta.1:173-181.

113. Urushihara Y, Suzuki T, Shimizu Y, Ohtaki M, Kuwahara Y, Suzuki M, Uno T, Fujita S, Saito A, Yamashiro H, et al. 2018. Haematological analysis of Japanese macaques (Macaca fuscata) in the area affected by the Fukushima Daiichi Nuclear Power Plant accident. Scientific reports.8:16748

114. Wang Q, Paunesku T, Woloschak G. 2010. Tissue and data archives from irradiation experiments conducted at Argonne National Laboratory over a period of four decades. Radiation and environmental biophysics.49:317-324.

115. Wegener K, Hasenohrl K, Wesch H. 1983. Recent results of the German Thorotrast study--pathoanatomical changes in animal experiments and comparison to human thorotrastosis. Health physics.44 Suppl 1:307-316. Epub 1983/01/01.

116. Wicherts JM, Bakker M, Molenaar D. 2011. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PloS one.6:e26828.

117. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data.3:160018.

118. Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. 2018. A design framework and exemplar metrics for FAIRness. Scientific data.5:180118. Epub 2018/06/27.

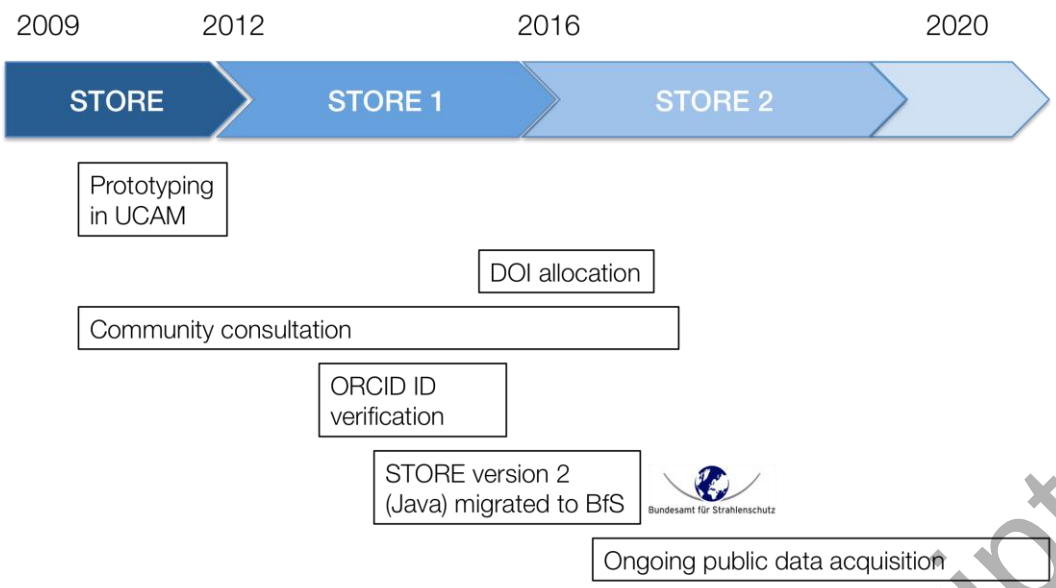
119. Yoshikane T, Yoshimura K. 2018. Dispersion characteristics of radioactive materials estimated by wind patterns. Scientific reports.8:9926.

120. Zander A, Paunesku T, Woloschak G. 2019. Radiation databases and archives - examples and comparisons. International journal of radiation biology.1-12. Epub 2019/01/25.

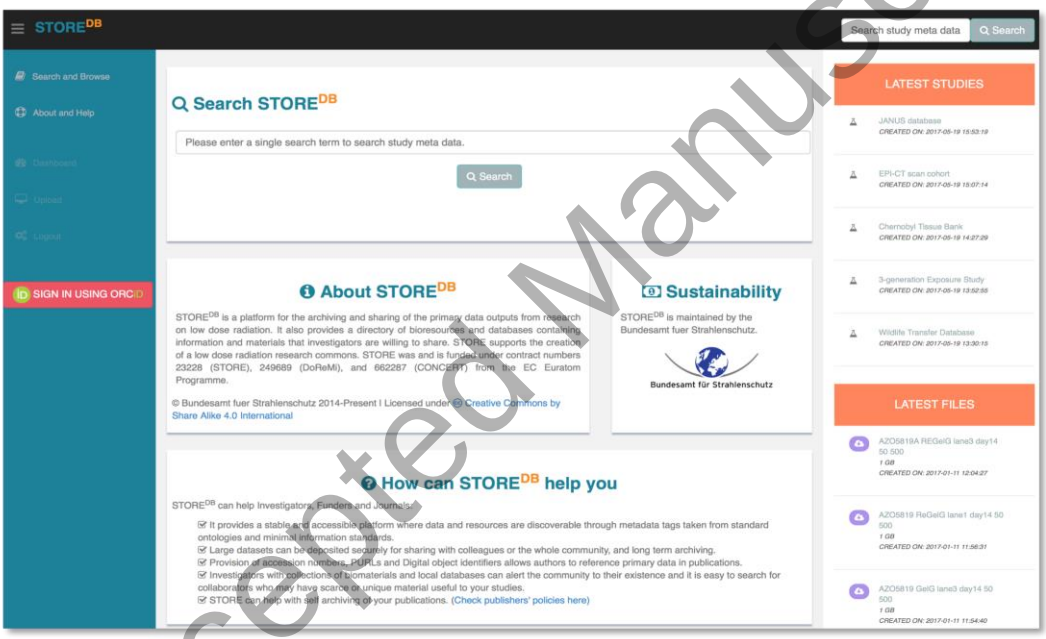
121. Zhao JZL, Mucaki EJ, Rogan PK. 2018. Predicting ionizing radiation exposure using biochemically-inspired genomic machine learning. F1000Research.7:233. Epub 2018/06/27.

Figure Legend

Figure 1. A; timeline for the triphasic development of the STORE database. Development of the STORE database and initial prototyping was carried out between 2009 and 2012. Community consultation was a critical part of Phases 1 and 2, with outreach to different radiobiological communities and detailed development of the data structure and user interfaces. Integration with the ORCID programme to allow users to be authenticated through their ORCID IDs coincided with movement of the physical database from Cambridge to the BfS in Neuherberg and the rewriting of the database backend using JAVA in order to be compliant with the BfS computing environment. At this point STORE DOIs were enabled and the database fulfilled criteria for stable identifiers with the identifiers.org project and recognised by re3data and FAIRsharing. Since 2016-17 the main activity of STORE has been the acquisition of data, both solicited datasets and community-driven uploads. **B**: A screenshot of the front page of the STORE database; http://www.storedb.org. STORE was and is funded under contract numbers 23228 (STORE), 249689 (DoReMi), and 662287 (CONCERT) from the EC Euratom Programme.



A



B

Table Legends

Table 1. Summary table of studies discussed. Studies are listed in the order in which they appear in the text and the most significant citation for each entry is listed in the table. Where there are no pblications of which we are aware a link is provided to the resource or to an online description. **Abbreviations for responsible Institutions.** Argonne; Argonne National Laboratories, USA, BfS; Bundesamt fuer Strahlenchutz, CIEMAT; Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, DKFZ; Deutsche Krebsforschungszentrum, HMGU; Helmholtz Centrum Muenchen, ICL; Imperial College, London, IES; Institute for Environmental Science, Rokkasho, Japan, IRME; National Research Institute of Radiation Medicine and Ecology of Kazakhstan, NURA; Northwestern University Radiation Archive, PHE; Public Health England, Chilton, UK, QST-NIRS; National Institute of Radiological Sciences, Japan, RERF; Radiation Effects Research Foundation, Japan, SUBI; South Urals Biophysics Institute, UCAM; University of Cambridge, UK, UKAE; UIAR; Ukraine Institute for Agricultural Radiology, UK Atomic Energy Authority, USTUR; United States Transuranium and Uranium Registries, URCRM; Urals Research Center for Radiation Medicine.

Dataset or database	Data Subject		Exposure type	Date collection started	Institute	Number of individuals	Purpose	Citation
	Biota	Non-Biota						
Lifespan Study (LSS)	human		external	1946	RERF	120, 000	Followup of Hiroshima and Nagasaki A-Bomb survivors	81
Mayak	human		mixed	1948	SUBI	12,585	Occupational exposure at the Mayak Plant	10
Techa River accident	human		internal	1949	SUBI	29, 000	Population followup after accidental release of radionuclides into the Techa river from the Mayak plant.	56
Semipalatinsk Registry	human		external	1949	IRME	300,000	Population exposures near the Semipalatinsk nuclear weapons testing site.	(7, 38)
Wismut Uranium miners study	human		mixed	1946	BfS	59, 000	Miners and workers occupational exposure during uranium production	58
German Thorotrast study	human		internal	1950s	DKFZ/BfS	2, 326	Patients exposed to thorium in clinical use of Thorotrast	37
Japanese Thorotrast study	human		internal	1950s	Tohoku University	436	Patients exposed to thorium in clinical use of Thorotrast	31
Kyshtym Accident	human		mixed	1957	URCRM	21, 400	Population exposure at the East Urals Radiactive Trace (EURT)	3
Chernobyl	human		mixed	1986	multiple	600, 000	Mainly liquidators exposed during Chernobyl clean-up	(21, 44)
	human		internal	1986	multiple	~30,000	Children from populations surrounding Chernobyl	(21, 44)
Fukushima TEPCO emergency workers	human		external	2011	RERF	5000	Fukushima emergency workers	43
INWORKS	human		external	1944	multiple	308, 297	Occupationaly exposed radiation workers from 3 countries	41
CEDR	human		mixed	1940s	US DOE	>1M	Health studies of DOE contract workers and environmental studies of areas surrounding DOE facilities	
EPI-CT	human		external	1977	multiple	948, 174	Children subjected to cranial CT scans	15
Argonne Janus Programme	mouse		external	1969	NURA/Northwestern University	>50, 000	Database of mice externally exposed experimentally during the JANUS programme	(39, 114)
Argonne Beagle Programme	Beagle		external	1961	NURA/Northwestern University	5, 000	Database of beagles externally and internally exposed	(39, 114)
IES institutional database	mouse		external	1996	IES	>10000	Mainly low-dose chronic external irradiation of mice	20
J-SHARE	mouse		external		QST-NIRS	13, 000	Mice and rats externally exposed to a range of radiation qualities	71
ERA	multiple		multiple	1960	UCAM/BfS	>400, 000	Agregative database of data from animal experiments conducted between the 1960s and the 1990s. Some human data also included.	16
FREDERICA	multiple		multiple	N/A	UK AE/CIEMAT		Data on the effects of radiation on non-human biota curated from the scientific literature	25
Wildlife transfer database	multiple	Soil, water, air, sediment	multiple	N/A	multiple		Parameter values for use in environmental radiological assessments to estimate the transfer of radioactivity to non-human biota	26
Radioecology Exchange	multiple	multiple	N/A	N/A	multiple		Aggregative data platform for multiple studies on environmental contamination	72
PROBA		Soil, water, sediment		1986	UIAR		Distribution of contamination from the Chernobyl exclusion zone	(53, 54)
Bioresource collection	**Biomaterial Source**							
Chernobyl Tissue Bank	human		internal	1966	hosted at ICL	4500	Biological samples from patients with thyroid tumors exposed as children or juveniles by fallout from the Chernobyl accident.	111
German Uranium Miners Bio- and Databank (Wismut)	human		internal		BfS/HMGU	463	Samples from workers with occupational exposure to Uranium production.	89
Adult health survey (LSS)	human		mixed	2013	RERF	15000	Biosamples from atomic bomb survivor followup studies	81
SUBI animal experiments	various		mixed		SUBI	6000	Multiple samples from a variety of species used in internal and external exposure	DOI: 10.20348/STOREDB/1056/1094
The Russian Radiobiological Human Tissue Repository (RHTR)	human		external	1951	SUBI	2000	Surgical tissues, blood and DNA from exposed workers at the Mayak facilities and local residents never occupationally exposed to ionizing radiation	61
J-SHARE	mice, rats		external		QST-NIRS	>15, 000	Frozen samples, paraffin blocks, and histopathological slides from rats and mice exposed to external radiation.	71
IES	mice		external	1996	IES	unknown	Tissue samples from mice subjected to long term low dose radiation	20
Sample bank of Fukushima animals	cows		mainly internal	2011	Tohoku University	1500	Domestic livestock collected from Fukushima evacuation zone organs were sampled, and stored as formalin fixed, paraffin embedded blocks or -80C	106
National Human Radiobiology Tissue Repository	human				USTUR	19000	Frozen and formalin-fixed tissue samples from exposed workers, plus historic samples	112
The Nagasaki Atomic Bomb Survivors' Tumor Tissue Bank	human		external	2008	Nagasaki University	600	Solid cancers and haemopoletic malignancies from atomic bomb survivors in Nagasaki	69
NURA	beagles		external		Northwestern University	1124	Paraffin embedded material from selected organs of experimentally irradiated beagles	(39, 114)
NURA	mice		external		Northwestern University	39, 000	Paraffin embedded material from selected organsof experimentally irradiated mice	(39, 114)
STAR Radioecology database	various	air, water , soil, building materials		2000	Radioecology Alliance		Project specific environmental datasets with wide aims and themes.	https://bit.ly/2MAu2s5

Table 2 Contents of the ERA database

Archives	Labs	Studies	Groups	Animals total	Animals with data
ERA	21	149	4,623	232,587	93,445
NRA	11	143	1,861	190,471	115,801
JRA	14	39	367	29,537	3,396
Total	46	331	6,851	452,595	212,642