Data Sets

Real Data Corpus

The Real Data Corpus (RDC) is a collection of disk images extracted from secondary storage devices that were acquired from second-hand markets around the world. In total, the RDC currently consists of 58 TiB of data contained in 3,127 disk images from 29 countries. A variety of devices are represented, including magnetic media and solid state storage from laptops, desktops, mobile phones, USB memory sticks, and other media. The dataset is hosted in the HPC infrastructure at the Naval Postgraduate School, as well as in AWS Govcloud.

Potential Uses

The Real Data Corpus is a one-of-a-kind scientific resource for:

- Developing and validating forensic and data recovery tools.
- Training students in forensics and data recovery
- Developing and validating document translation software.
- Exploring and characterizing real-world computing practices, configuration choices, and option settings.
- Studying the storage allocation strategies of file systems under real-world conditions

The RDC has been cited in over 60 articles. See our current list here.

Current Contents

The following countries are represented:

Data By Country

Country Code	Country Name	# Images	Size (readable)	Size (bytes)
8.03 TiB	United Arab Emirates	89	8.03 TiB	8,825,258,946,193
1.52 TiB	Austria	44	1.52 TiB	1,674,638,465,645
574.73 MiB	Bosnia	7	574.73 MiB	602,644,365
1.56 TiB	Bangladesh	59	1.56 TiB	1,710,737,896,119
1.38 TiB	Bahamas	34	1.38 TiB	1,521,076,921,944
893.59 GiB	Canada	54	893.59 GiB	959,485,134,238
1.61 GiB	Switzerland	2	1.61 GiB	1,727,236,374
561.15 GiB	China	746	561.15 GiB	602,527,863,126
1.38 TiB	Czech Republic	24	1.38 TiB	1,521,964,790,057
636.61 GiB	Germany	41	636.61 GiB	683,551,846,923
53.75 GiB	Egypt	7	53.75 GiB	57,710,165,396
622.91 GiB	Ghana	21	622.91 GiB	668,842,288,279
3.04 GiB	Greece	7	3.04 GiB	3,267,501,589
145.92 GiB	Hong Kong	8	145.92 GiB	156,677,292,656
510.46 GiB	Hungary	22	510.46 GiB	548,097,899,391
7.5 TiB	Israel	300	7.5 TiB	8,246,871,569,750
10.99 TiB	India	669	10.99 TiB	12,078,854,291,887
29.58 GiB	Japan	4	29.58 GiB	31,760,575,283
108.54 GiB	Morocco	11	108.54 GiB	116,547,412,932
403.45 GiB	Mexico	171	403.45 GiB	433,196,045,674
1.86 TiB	Malaysia	78	1.86 TiB	2,043,906,920,751
204.38 GiB	Panama	17	204.38 GiB	219,454,669,389
3.44 TiB	Pakistan	88	3.44 TiB	3,784,807,218,108

1.07 TiB	Palestine	139	1.07 TiB	1,174,201,653,174
818.9 GiB	Serbia	24	818.9 GiB	879,290,824,361
5.9 TiB	Singapore	238	5.9 TiB	6,491,155,492,690
6.99 TiB	Thailand	188	6.99 TiB	7,681,881,741,459
484.83 GiB	Turkey	10	484.83 GiB	520,583,203,000
850.58 GiB	United Kingdom	26	850.58 GiB	913,307,005,195
57.74 TiB	All	3128	57.74 TiB	63,487,653,727,660

Access and Availability

Please contact us if you would like access to the Real Data Corpus. In general, due to privacy concerns, we do not release copies of the data to private individuals. However, depending on the requirements of the project, we may be able to offer access through one of two methods:

1. *Mediated Access.* Researchers submit source code, build instructions, and detailed instructions for running their experiment. We return sanitized results. This is the most expedient option in cases where the desired experiment does not involve human subjects research.
2. *Direct Access.* Researchers create virtual machines on Amazon GovCloud, and these machines are granted access to the dataset. Because this method may involve direct contact with sensitive data, it involves additional review.

Please be aware that due to limited staff we cannot always accommodate all requests. Efforts are underway to develop infrastructure that will allow us to meet a wider range of research requirements without unduly increasing privacy risks.

IRB Required for Research

The National Research Act[2] (NRA) of 1974 and the Common Rule,[3] govern all federally funded research in the United States that is performed with human beings as experimental subjects. Because portions of the Real Data Corpus were funded by the US Government, this legal framework must be followed in research involving the Real Data Corpus. The Common Rule creates a four-part test that determines whether or not proposed activity must be reviewed by an IRB. Specifically, IRB approval is required if:

1. . The activity constitutes scientific “research,” a term that the Common Rule broadly defines as “a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.”[4]
2. . The research must be federally funded.[5]
3. . The research must involve human subjects, which the Common Rule defines as “a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information.”[6]
4. . The research is not “exempt” under the regulations.[7] The Common Rule exempts research involving “existing data, documents, [and] records…” provided that the data set is either “publicly available” or that the subjects “cannot be identified, directly or through identifiers linked to the subjects”(§46.101(b)(4)).

Research involving the Real Data Corpus is not exempt under the Common Rule because the RDC is not publicly available and in many cases it is possible to identify individuals whose data are in the collection. Furthermore, the majority of the subjects included in the Real Data Corpus have not provided consent to have their data used for research. Mitigating factors allowing the use of this data is the fact that the data was lawfully obtained, research involving this data is “minimal risk” (provided that the data is properly protected and personally identifiable information inside the RDC is kept confidential), the fact that there is substantial public benefit in using the RDC for research into computer forensics and computer security, and the fact that there is no practical alternative to using this data. Even if research involving the RDC were exempt, most US universities do not allow experiments to make their own determination of exemption. Instead, these institutions require that the experimenter submit an application for exempt research to the IRB. To date no IRB has blocked the approval of research that involves the RDC. In order to submit an application to an IRB it is necessary for all experimenters who will make use of the human subject data to take the appropriate human subject training proscribed by their institution. Most institutions prohibit students from filing applications directly, and instead require that an application be filed by a researcher or professor that can be considered a “principal investigator” for external funding. As a result, any proposed use of the RDC in research requires that an IRB application be filed with the host institution and with the Naval Postgraduate School. A copy of both the application and the approval from both the host institution and NPS must be provided prior to access being granted. The application must clearly state:

- The proposed research that is to be done.
- Why it is necessary to use the RDC; why simulated or realistic data cannot be used as an alternative.
- What measures will be used to protect the data in the RDC.
- What measures will be used to prevent the publication of personally identifiable information in any research products.

Please provide us with your IRB application prior to submitting it to your IRB! We can review the application and let you know if it is consistent with the IRB approval that we have already approved, or if we will need to apply for additional IRB approval. Sample applications are available upon request.

Contact Information

For more information or if you're interested in access to the Real Data Corpus, please contract:

Brittany Ramsey - Research Associate

blramsey@nps.edu (831) 656-2014