RESEARCH RESOURCE CENTER

The Research Resource Center (RRC) is envisioned to be a core component of RERF's infrastructure. The successful implementation and execution of the RRC is a necessary component to advance RERF's strategic plans.

The RRC's mission is 3-fold. The mission includes:
1. Protect, index, and integrate RERF's research assets. These include data, biosample inventories, paper records, artifacts, manuscripts, datasets and programming scripts, as well as other historically important articles. Access to research data will be made through a web portal with clear accessibility rules that protect the privacy of our subjects.
2. Enhance RERF's ability to perform research by integrating all data and biosample inventories. Tools for data visualization, data assembly, and analysis will simplify and standardize access, and facilitate research.
3. Provide an administrative framework to facilitate internal research as well as facilitate collaborative research projects, including contracts and grants.
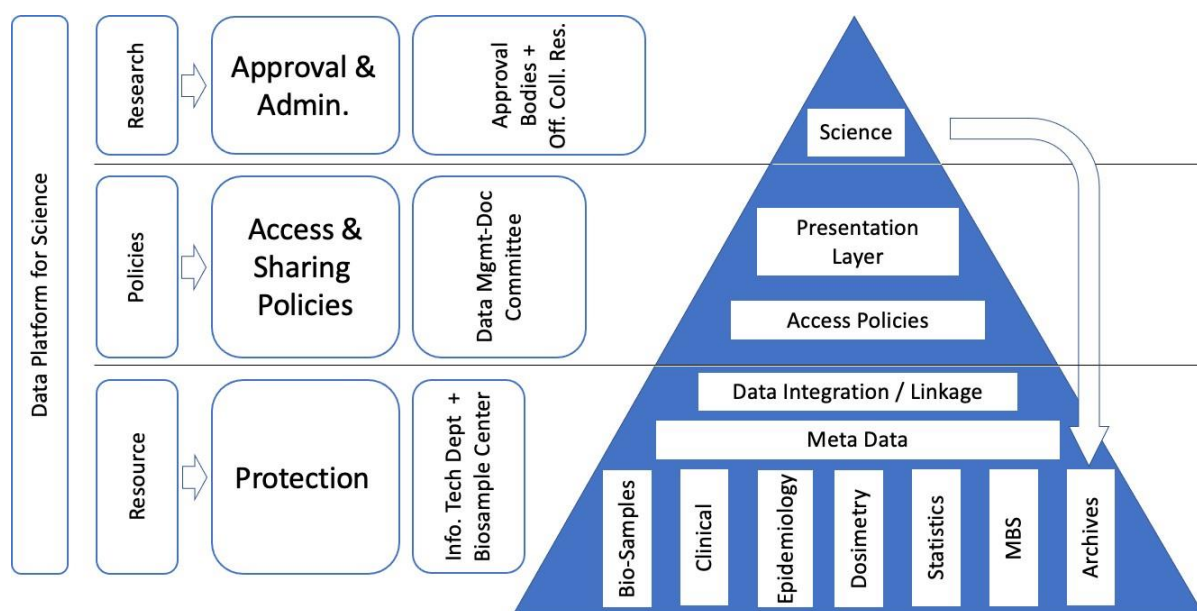

Figure 1. Functional structure of RERF's Research Resource Center

Referring to Figure 1, the bottom layer of the pyramid is for protection. This will include a framework for digital data storage, integration, and visualization. Beyond digital data, it will also include a Content Management System (CMS) to catalog and index all digital scans of RERF materials and inventories of historical artifacts. Copies of data and scans will be housed centrally (Mission #1). All materials will have information describing each entry ("metadata") so that the entries can be located, be appropriately associated with related materials, and be searchable. The second layer is the "Policy and Access" layer. Using institutional rules and modern authentication methods, only those materials to which a person is authorized will be viewable. This layer will include the Presentation tools so that researchers (both internal and external) can peruse RERF data catalogs. Authorized users can also see research proposals and

manuscripts based on those data and visualize/query the data to design new research (Mission #2). The top layer is the Science Layer. Here, authorized research can be performed. If the research is collaborative, the "Office of Research Support" (ORS) can facilitate the research by executing contracts and assisting with procedures to distribute data and process biosample requests (Mission #3). At the conclusion of a study, all the materials associated with the study (data, assay results, analysis scripts, manuscripts, etc.) will be placed into the Archives with appropriate metadata and linkages so that those products can be reused by future researchers.

Personnel and functional units to develop/manage the RRC have been tentatively set by the Research Resource Center Operating Committee. This ad hoc committee was initiated in November 2021 and had its first meeting in December 2021. It is led by Dr. Eric Grant as chair and Dr. Tanabe as co-chair and includes all department chiefs. This committee supersedes a previous ad hoc committee entitled the "Preparatory Committee for the Establishment of the Research Resource Center", which was established in December 2019. See Figure 2 illustrating the functional units and reporting lines of the RRC.

Three primary development needs

To realize the RRC with the functionality outlined in Figure 1, there are three broad development areas that need to be completed. These three areas are a "data framework", a "content management system", and the Office of Research Support (ORS). The functionality and the development areas have been carefully thought through by an RERF "Technical Team" that started biweekly meetings in 2017. The team includes senior members of the ITD and Dr. Grant.

Data Management Framework. The "data management framework" will be the underlying computer environment for RERF data. It must be able to store, annotate, and integrate data. Ideally, it would be flexible so that new data can be easily added to the system using FAIR principles. Ideally, annotation would be possible by the researcher who created the data (rather than forcing computer-savvy administrators to perform this work). The system would support both English and Japanese, could be installed on local servers or in a cloud environment, and would allow remote researchers to run analyses in situ so that the data don't "leave" the institute. There would need to be a hierarchical and highly secure authentication system so only authorized personnel could access the data. It would also allow "data visualization applications" to interact with the underlying data so that it could be visualized and queried for data exploration.

A pilot for such a system (The Gen3 Data Commons, developed by the University of Chicago) was run for over a year in 2020-2021. Unfortunately, the pilot was not successful for a variety of reasons discussed later in this chapter. RERF is currently searching for a suitable data management framework. This is one of the biggest current challenges facing the RRC.

A pilot project with data visualization software is currently underway and will be demonstrated for the SAC.

Content Management System. The Content Management System (CMS) is a system for storing all paper/text-based assets within RERF. A flexible and strong security system that

allows access to RERF materials while strictly adhering to RERF's authorized access policies is required.

For more general materials (i.e. lists of publications, etc.), RERF currently has a number of databases for viewing various lists (Research Protocols, Technical Reports, etc.) but they are often separate databases (i.e., multiple sites need to be interrogated), and there is no full-text search ability anywhere within RERF. These issues (no centralized access, no full-text search) are major shortcomings in RERF's current computer environment. RERF needs a CMS that can house scans of paper records within RERF, all RERF pdf files, lab notebooks, questionnaires, etc. The system needs to be bilingual, intuitive, and allow individuals to upload locally held assets for preservation and re-use.

The RRC Technical Team has identified a CMS system and has hired an outside contracting firm specializing in this CMS to develop a pilot project. A demonstration of this system is planned for the SAC.

Office of Research Support. Performing research at RERF is unnecessarily slowed by administrative bottlenecks. All work is performed on paper and many approval forms overlap or conflict with other approval forms. Further, there is no "presence" on RERF's external home page for outside researchers to peruse RERF holdings or initiate conversations to explore collaborative research projects. The Office of Research Support (ORS) would try to solve each of these problems. We envision a thorough review of RERF's application forms and redesigning them to avoid overlap. Further, the entire process should be moved to an online system. This would be developed using an Enterprise Resource Planning (ERP) package that would allow the creation of customized workflows. A new person with bilingual abilities will need to be hired.

A tentative budget has been developed that includes personnel, contractor support, and software procurement. Please see Figure 3.
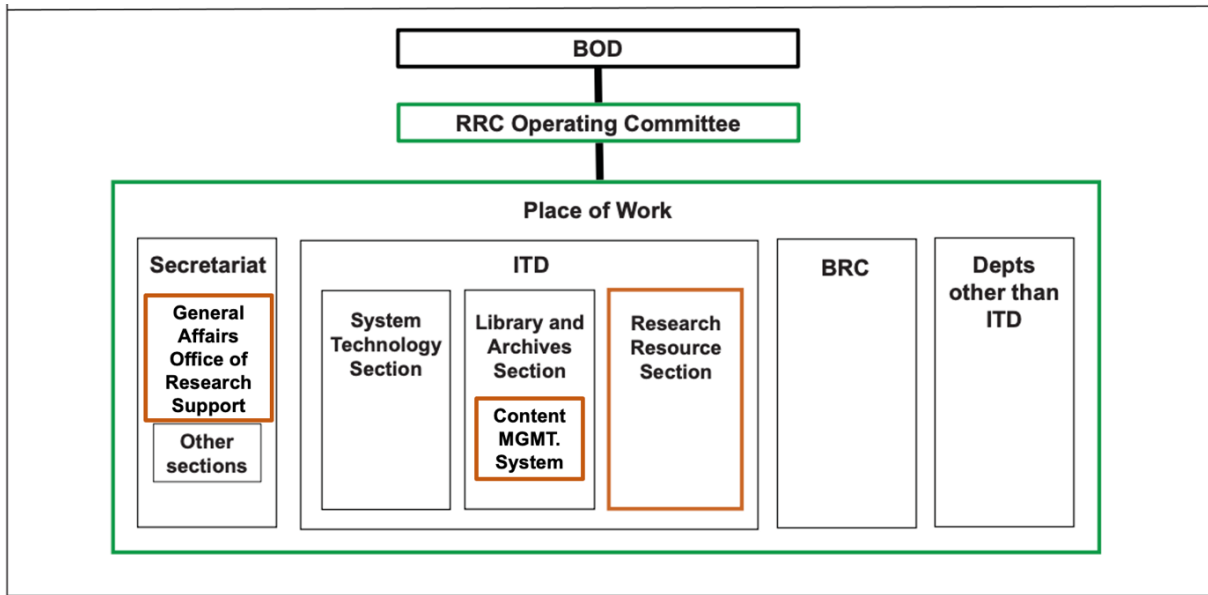
Achievements

Figure 2. Structure of the RRC. This is the first of a two-step process where sections are created within existing units at RERF. The second step will be to establish the RRC as a "stand alone" center with its own personnel and physical location.

RRC Budget Estimates ($1000s)

	FY2022	FY2023	FY2024	FY2025	FY2026
Personnel					
Contract Programmer	100	100			
ITD Programmer		50	50	50	50
Off. Res. Support				50	50
Software + External Contractors					
Content Management Sys.	50	10	10	10	10
Data Visualization S/W	20	20	20	20	20
Cloud Computing	0	100	100	100	200
Data integration					
Pilot	50	50	50		
Full-scale implementation				50	10
TOTAL (by year):	220	330	230	280	340
TOTAL (5-years):	$1,400,000				

Note: Assumes ERP system (online procedures S/W) paid by Secretariat; No budget for large-scale scanning ($10M)

Figure 3. Tentative budget estimates for developing the RRC (5-years). Note that this does not cover the cost of volume scanning, estimated to be roughly $10M for all RERF paper sheets (estimated to be about 20M sheets).

Achievements

FY2021 Departmental Achievements

The primary accomplishment of the Research Resource Center in 2021 was the re-design of its administrative structure.

Initially, plans for the RRC were led by the "Preparatory Committee for the Establishment of the Research Resource Center", established in December 2019, with Drs. Niwa and Ullrich as co-chairs. This ad hoc committee was established to inform and garner the support of all departments. Within the ad hoc committee, a "Drafting Guidelines" subcommittee was responsible for determining the position of the RRC within the institution, including reporting lines, leadership, budget, personnel, and internal structure of the RRC. Its 5th meeting was held in 2021, by which time several projects involving inventory and scanning preparations had been completed. The ad hoc committee was disbanded in November 2021 with the establishment of a new ad hoc committee "Research Resource Center Operating Committee" (RRCOC), chaired by Dr. Eric Grant with Dr. Tanabe as the vice-chair. The first of the RRCOC was held in December 2021.

The RRC was initially envisioned as a stand-alone "Center" within RERF with its own physical presence and personnel. This is still the long-run plan. However, to speed development, a "two-step" process has been embraced to initially establish several new sections within existing units at RERF and then later bring the groups together as the RRC. As described in the "Overview" section, there are three primary development areas needed to make the RRC functional. They include 1) The installation and operation of a data framework for data integration, data tagging, data visualization, and dataset assembly. 2) A Content Management System that will centralize all written materials within RERF and provide full-text search (bi-lingual). 3) An Office of Research Support that will be tasked with streamlining all application/approval procedures and moving these procedures to an online system, as well as establishing an online presence for the RRC. To accomplish these tasks, the RRCOC has approved the creation (or repurposing) of three sections. They are:

- **Research Resource Section.** This new section (the "RRS") will be a technology-based section located within the department of Information Technology and will be led by the Department Chief, Dr. Ono. The Section will be tasked overseeing the installation of a "data management framework" that will provide a flexible system for centralizing RERF data, including tagging the data, visualizing the data, and assembling new datasets for research. The section will start by hiring a senior contract programmer to lead the RRC programming efforts. In FY2023, a new programmer will be hired into the Section. The Section will further be bolstered with two (existing) programmers that will be concurrently assigned from ITD's Systems Technology Section to assist with data integration and development tasks.
- **Library and Archives Section.** This section already exists within RERF but will be given the additional assignment of identifying, uploading, and cataloging manuscripts, research protocols, technical reports, etc. into the new Content Management System (CMS). They will also work with outside vendors to customize the CMS for RERF's needs. The Section reports directly to Dr. Ono.
- **Office of Research Support.** This section will be newly established within RERF's Secretariat. The Secretariat is will-versed with regulations and policies enforcement. A new person (with bi-lingual skills) will be hired and will report to the Chief of the Secretariat. This person will oversee an overhaul of RERF research procedures to reduce

administrative burdens. This will be accomplished via streamline application processes and approval channels, moving all procedures to an online system, establishing a web presence to both internal and external researchers to facilitate both internal and collaborative research. Its placement within the Secretariat also ensures that it will have close communication with the Accounting Section to support Grant applications and other governmental filing requirements.

In addition to the RRCOC setting up the future structure and personnel requirements of the RRC, several projects have been underway.

- **Five-year timeline for establishing the RRC.** Dr. Grant and Dr. Ono have worked together to create a timeline for technical developments within the RRC. See Figure 1. Highlights of the timeline include several infrastructure improvements at RERF required for technical advancement (e.g., a single sign-on system for all RERF resources including multi-factor authentication), and the introduction of an Enterprise Resource Planning (ERP) system to allow the automation/streamlining of a number of processes that are currently handled in a manual, paper-based manner within RERF. Other highlights include the timelines for establishing new sections to handle the tasks discussed above. The RRC should be in full operation by the end of FY2025 with an external on-line presence established by the end of FY2024.

- **Scanning Center Pilot Project Initiated.** Under the leadership of Dr. Ono, the ITD has reconfigured a number of RERF multi-purpose copy machines for use in RERF's "Scanning Center". The Scanning Center is a facility by which individual researchers or Sections can begin to move personal notes and papers to a digital format. After logging in to the scanning machine, all scans are sent through an Optical Character Recognition feature and the newly created pdf file is centrally located under the creator's name. These files can then be re-arranged for storage. Eventually, these files will be automatically moved to RERF's CMS system for tagging and will be assigned an appropriate (and modifiable) access profile (which can access the document).

- **Content Management System Pilot.** RERF identified Open Source software for content management and then located the top programming firm in Japan specializing in the software. We then contracted that firm to develop a pilot system for RERF's content management. This is progressing but we have learned that RERF's design specifications are quite challenging. They include: A hierarchical access system that is compatible with RERF's single sign-on system, fully multi-lingual, complex relational links between various materials, flexible, and customizable. We are hopeful that this system will meet our needs but this should still be considered a "pilot" and success is not assured.

- **Data Visualization Pilot.** RERF has purchased a limited number of "seats" for "Tableau" data visualization software. This software is a leader in the field of data visualization and can be used as a 'front end' to explore RERF data. As with any complex software, there is a significant learning curve. The software looks promising but success will require personnel be dedicated to learning and developing the package for use at RERF. Note: a demonstration is scheduled for the SAC meeting.

- **MEXT Grant submitted.** A MEXT grant in collaboration with Hiroshima University's

Achievements

Research Institute for Radiation Biology and Medicine (RIRBM) was submitted. The grant passed its first funding hurdle and the process moved to the second, interview, stage but was ultimately not funded. The collaborative application between RERF and RIRBM was to establish a public-facing web-based archive system to give the public access to jointly held historical materials of interest to the public and social scientists. The design was to have federated databases (each institute can independently control which materials are provided) with a single interface for search and presentation. Knowledge gained from the design of this system was intended to be leveraged for inclusion in the creation/design of the RRC. The grant will be re-applied for in FY2022.

- **Digitization of deteriorating microfilms.** Several hundred microfilms located in the Master File Section and the Clinical Studies Section were deteriorating due to poor storage conditions over many decades. These films were identified and all have been digitized by an external vendor. Recovery was generally excellent, with only 4-5 reels being deteriorated to the point that they could not be digitized using automated machinery. They will be flagged and stored. If deemed absolutely required, they could be recovered on a slide-by-slide (manual) basis.

- **"Gen3" Data Commons Pilot Project terminated.** Following a site visit from Dr. Robert Grossman from the University of Chicago, RERF decided to attempt a pilot project using the Gen3 Data Framework, developed by the University of Chicago and used by the US National Cancer Institute. Gen3 is an open-source software platform for large scale data storage and analysis. RERF had no expertise or available personnel when this pilot was discussed. Through contacts, Dr. Grant recruited a computer scientist working on his PhD in Epidemiology from the University of Texas Medical Center to perform an internship with RERF. The initial proposal was for the programmer to visit the University of Chicago for several weeks of training in the spring 2020 and then visit RERF through the summer to work full time on the Gen3 installation in Hiroshima. However, with the COVID pandemic, all site visits had to be canceled. To his credit, the computer scientist met with Dr. Grant for weekly teleconferences throughout the summer and into December 2020. Unfortunately, the pilot project was terminated. This is mostly a confluence of bad circumstances rather than a fault with any individual. Compounding factors included the nascent state of the Gen3 system with incomplete documentation, the Gen3 engineering staff being overwhelmed by their day-to-day activities and unable to give sufficient time to the RERF installation, and a graduate student trying to perform this project remotely. The RRC Technical team learned that large, technical systems require a technical team to install and manage (cannot be a "shoestring" effort by a minimal staff). Also, the state of the development of the Data Framework should also be carefully evaluated. The more mature the state, the easier the installation. Systems that are new and rapidly evolving require engineers that are intimately tied to the development teams.

- **Bi-weekly Technical Team meetings.** Dr. Grant, ITD leadership, and members from Systems section and Library & Archives section initiated bi-weekly meetings in 2019 to scope the RRC into different tasks and begin planning for implementation. These discussions range from high-level conceptual discussions to specific features. Significant time and research were devoted to technical needs and identifying groups/software platforms that are currently implementing our desired features. This group has focused on technical issues. These meetings continue outside the scope of the formal committees.

Achievements and Future Plans

TASKS	Jan-22	Feb-22	Mar-22	Apr-22	May-22	Jun-22	Jul-22	Aug-22	Sep-22	Oct-22	Nov-22	Dec-22	Jan-23	Feb-23	Mar-23	FY2023 first half-year	FY2023 latter period	FY2024 first half-year	FY2024 latter period	FY2025 first half-year	FY2025 latter period	FY2026 first half-year	FY2026 latter period
4 Above new people employed period										███	███	███	███	███	███	███	███	███					
5 Establishment of RRS				███																	★ RRC Online (external)		
6 Activity of hiring new people (2024 new graduate)								███	███														
7 Start employment new graduate staff																	███	███	███	███	███	███	███
																OJT (On the Job Training) period							
																for new graduate staff						
DATA INTEGRATION																							
1 Agreement Data Integration	███	███	███	███																			
2 Biosample inventories database to Ondo DB	███	███	███	███																			
6 Investigate Siloed data	███	███	███	███	███	███	███	███	███	███													
3 Data visualization pilot project / demonstration		███	███	███	███																		
4 Data visualization tool operation start						███																	
5 Identify potential data frameworks for RERF (e.g. Gen3, IBM Cloud pack4)							███	███	███	███	███	███	███	███	███								
7 Evaluate data framework #1 (pilot project#1)								███	███	███	███												
8 Evaluate data framework #1 (pilot project#2)											███	███	███										
9 Evaluate data framework #1 (pilot project#3)														███	███								
10 Specification formulation of the data framework system																███	███						
11 Procurement of the data framework system																	███	███					
12 Construction of the system (by a vender)																			███	███			
13 Metadata tagging for all data moved to new framework																					███	███	███
14 Considering Cloud Based System (include pilot project)									███	███	███	███	███	███	███	███	███	███					
15 Regarding procurement of the Cloud Based System																		███	███	███	███		
16 Migration to the Cloud Based System																							███
CONTENT MANAGEMENT SYSTEM																							
1 Pilot Project (Create Demo site)	███	███	███	███																			
2 Specification formulation of the system				███	███	███																	
3 Procurement of the system						███	███	███															
4 Construction of the system (by a vender)								███	███	███													
5 Digitized of Data (Academic material)				███	███	███	███	███	███	███													
6 Digitized of Data (Historic material)										███	███	███	███										
7 Migration of all digitized data to CMS													███	███									
8 Initiate Scancenter				███	███	███																	
9 Link Scancenter to CMS						███	███	███															
10 Create a Metadata to the materials																			███	███	███		
11 Starting CMS operation																					███	███	███
OFFICE OF RESEARCH SUPPORT																							
1 Identify all procedures regarding research						███	███	███	███	███	███	███	███										
2 Specification formulation of the system for applications/approvals (ERP)														███	███								
3 Procurement of the system (ERP)																███							
4 Construct of the System (ERP)																	███						
5 Create online (external) web presence for external researchers																	███	███					
6 Activity of hiring new people (2024 new graduate for ORS)																	███						
7 Start employment new graduate staff (for ORS)																					███	███	███

RESEARCH RESOURCE CENTER

Achievements

While not occurring in 2020, development efforts for the RRC started in 2017. Here is a brief list of other activities/achievements that occurred prior to 2020.

1. Visited data and archiving specialists from around the world. As the ITD staff do not have existing expertise for data integration or archiving, it is necessary to contact experts from various institutes involved in large-scale projects to understand options and development strategies. Face-to-face meetings were held with the following experts/institutes:

   - Dave Thompson/Digital Curator, Wellcome Library
   - Stephen Chanock/ US NCI, Director of Cancer Epidemiology and Genetics
   - Anthony Philippakis/Broad Institute, Chief Data Officer
   - James Cuff/Harvard University, Assistant Dean for Research Computing
   - Brian White/Harvard Univ., Manager of Research Computing IT Operations
   - John Quackenbush/Dana-Farber, Biostatistics and Computational Biology
   - Meir Stampfer/Harvard SPH, PI Nurses Health Study
   - James Lacey/California Teachers Study, Director Health Analytics
   - Tom Murphy, UMich. ICPSR, Director of Computer and Network Svcs.
   - Robert Grossman/UChicago, Chief Research Informatics Officer, Division of Biological Sciences
   - Nobumichi Ariga/Japan National Museum of Nature and Science, Department of Science and Engineering
   - Shigeo Sugimoto/University of Tsukuba, Library, Information and Media Science

2. Site visit by Dr. James Cuff (Harvard), May 2017

   Dr. James Cuff, Assistant Dean for Research Computing, Harvard University came to RERF for a 3 day visit to evaluate RERF's research needs and current infrastructure. Cuff worked with ITD and RERF leadership to outline a "White Paper" in an effort to scope the problem and break it down into logical steps.

3. Site visit by Dr. Robert Grossman (University of Chicago), January 2019

   As a follow-up to a visit to the University of Chicago where discussions or RERF's needs and whether the Gen3 system may be a good solution for RERF's data integration needs, Dr. Robert Grossman, Chief Research Informatics Officer, Division of Biological Sciences visited RERF for two days. He gave a talk ("Building a Data Center with Gen3") and met with visiting archivists, our ITD staff, and RERF senior management. At the end of the meeting, it was decided that a pilot program using the Gen3 system should be implemented at RERF. The pilot ran through January 2021 and was terminated. The Gen3 system was complex (and everchanging due to its "beta" software cycle). This coupled with a part-time student working on the system was not workable. In an exit interview with Dr. Grossman, he agreed that an "offsite" installation would be very challenging as the system was best run on their servers with their

engineers.

4. Site visit by Dr. Christy Henshaw (Wellcome Trust), March 2019

   Dr. Henshaw, Digital Production Manager, delivered a talk at RERF entitled "Digital Engagement at Wellcome Collection" in which she discussed the strategy of the Wellcome's online presence as well as many of the technical details in their operations. Dr. Henshaw also met with our archivists and other local archivists invited to attend her lecture. Many of the tools that Wellcome uses are open source and we discussed possible collaborative projects that may assist us in our RRC development.

5. Purchased servers to support the RRC

   After the pilot project was agreed upon in the January meeting, discussions of technical requirements began. The Gen3 team recommended a minimum of two high-spec servers. RERF decided to acquire 3 servers (two for Gen3 and one to support the web server with search and archives). In summer, the Supply Section commenced an open bidding process. The computers were purchased and delivered in September 2019.

6. Site Visit by Robert Grossman and two engineers

   In November 2019, Dr. Grossman and two engineers from the University of Chicago came to install the pilot version of the Gen3 Data Commons on the newly acquired servers. Dr. Grossman stayed two days and the engineers stayed for the full week. Our top systems engineer worked together with the engineers for the entire week to gain an understanding of the installation tools used and the network setup required. At the end of the week, an initial installation had been completed. This system also required RERF to adopt a modern authentication system using a 3rd-party authentication system with certificates (Microsoft Azure infrastructure).

7. Created two secure computing rooms with video surveillance and biometric entry for collaborative human genome research

   RERF is introducing human genomic data for which we do not have sufficient in-house computing power. To perform this type of research, access to super computers is needed (our initial target is the Tohoku Medical Megabank supercomputer). To use this computer, an approved physical space and computer infrastructure was required. Using construction vendors and specifications, two rooms have now been built at RERF for this specialized function.

8. Staff members attended conferences or lectures on Open Library Services, Open Data Initiatives within Japan, Conference on Open Repositories, lectures on Professional development

Achievements

9. Completion of White Paper

10. Completion of efforts by the "Digitization Subcommittee" (10 persons under Dr. Ono and Ms. Marumo/ITD) and the "Data Inventory Subcommittee" (13 persons under the direction of Dr. Sakata/Epidemiology).

    These two committees have completed inventorying all materials within RERF and identifying those that will need digitization. After numerous meetings and interviews with all departments, the Inventory Subcommittee identified and organized lists of 20 different types of materials (digital files, notebooks, maps, coding manuals, paper data cards, lab photos, etc.). This represents millions of paper sheets, digital records, filing cabinets full of folders, boxes, etc. The Digitization Subcommittee worked together with the Inventory Subcommittee to identify materials that require scanning.

11. Medical chart scanning pilot project completed.

    Via a contract with an external vendor, a pilot project was performed to scope the effort required to scan the paper medical charts stored in the Hiroshima Department of Clinical Studies. After a three-week trial period testing various work-flow methods, overhead vs. flatbed scanners, methods of dealing with various fasteners in varying states of decay, and handling odd-sized paper, a report was written and delivered to RERF. The pilot project estimated that there were 10,000,000 sheets of paper contained in the medical charts. The vendors concluded that digitizing all of the materials in the medical charts would cost roughly $5M dollars. A 25-member team could complete the job in 3 years. A smaller team could do it for the same cost, but it would take a longer period of time. It is estimated that Nagasaki materials would add roughly 40% to the total costs.

    In addition to paper resources identified in the Clinical Studies departments, it is estimated that another 8-10M sheet exist in other departments. This raises the total sheet count to roughly 20M. At roughly $0.50 per sheet, the total costs for external (contract-based) digitization of all RERF paper assets is estimated to be $10,000,000.