Advancing Scholarship through the Research Data Program

MICHAEL HEMMENT

Harvard Business School, Cambridge, MA

mhemment@hbs.edu

KATHERINE MCNEILL

Harvard Business School, Cambridge, MA

ORCID ID <https://orcid.org/0000-0003-2865-3751>

kmcneill@hbs.edu

Abstract Over the past five years, Baker Library at the Harvard Business School has developed a more formal program to facilitate the use of data throughout the lifecycle of business research. Driven by user needs and institutional data retention requirements, the Research Data Program (RDP) has established and brought together services in the realms of data acquisition and discovery, data curation and management, advice on research methods, and data sharing and archiving. These efforts are in line with a field-wide trend of librarian-led efforts to manage, share, and preserve institutional research data. One initiative of Baker's RDP, the Research Datasets Tool, is a searchable, secure discovery platform designed to enable the institution to log and find datasets purchased by individual faculty. Rollout of the Research Datasets Tool illustrates the benefits and challenges common to efforts to track institutional assets: increased ease of access and maximization of University resources are tempered by the burden of user education and the difficulty of integrating various information discovery systems. Over the next several years, Baker will continue to develop its existing research data tools and services, explore new areas of service, and work to better integrate its services with the larger university community.

Keywords *data curation, research data management, data preservation, discovery tools*

Introduction

Over the past five years, Baker Library (Baker) at the Harvard Business School has been building a more formalized, extensive program to facilitate the use of data throughout the lifecycle of business research. The Research Data Program utilizes staff across Baker to provide key services in partnership with other staff throughout the School and University. This article briefly outlines the evolution of the program, describes the main service portfolio, highlights the development of one particular service, shares challenges and lessons learned, and discusses future directions.

Researchers in the social sciences, including those in business, have a long history of working with data, utilizing data they have personally collected or, more commonly, that collected by other researchers, government, and business. Academic libraries, including those focused on business, have a comparable history of supporting this research, most traditionally in providing reference services to support data discovery. Additionally, reference librarians focused on supporting the use of numeric data in the social sciences have long been a fixture in academic libraries. Building on this history, recent years have seen a new wave of academic library data services designed to support the full research data life

cycle in all disciplines, including not only access to licensed data but also the management and sharing of university-produced research data. These services have been driven by a number of factors, including

- the research potential seen in the re-use of existing data;
- an opportunity for libraries to provide services to researchers to support the researchers' role as information producers, not simply information consumers (in line with their support for open access publishing);
- funder and publisher data sharing requirements; and
- a desire for HBS to better manage its research assets.

An extensive set of literature documents the historical development of university data services. Two publications that cover this topic extensively are *IASSIST Quarterly*¹ and the *International Journal of Digital Curation*². These two publications over the past decade document the development of a variety of different types of data and research data management (RDM) services in academic libraries. A sampling of articles profiles institutions such as Cornell University (Steinhart, 2009), The Georgia Institute of Technology (Walters, 2009), Leiden University (Schoots, Sesink, Verhaar, & Frederiks, 2017), Penn State University Libraries (Hswe et al., 2011), the University of Porto (Ribeiro & Fernandes, 2011), as well as recurring articles from the University of Edinburgh (Rice & Haywood, 2011; Rice et al., 2013) and Oxford University (Wilson, Martinez-Urbe, Fraser, & Jeffreys, 2011; Rumsey & Jefferies, 2013; Wilson & Jeffreys, 2013), which illustrate the evolution of their services over time. Thus the data services developing at the Harvard Business School continue a rich and active history and illustrate the current momentum in the field.

Context and Service Development at Harvard Business School

Baker Library serves the Harvard Business School (HBS) at Harvard University. Founded in 1908, the Harvard Graduate School of Business Administration (later renamed Harvard Business School) established the world's first MBA program with 15 faculty and 33 regular students. The school now runs a two-year MBA program (1,850 students), a robust doctoral program (140 students), an executive education program (10,000 participants annually), and 10 Global Research Centers, all under the auspices of more than 230 faculty. Baker's executive director reports to the dean of the school and oversees activities such as

- support for teaching and learning in educational programs,
- acquisition and management of contemporary and historical collections,
- participation in faculty and doctoral research production and dissemination,
- information services for alumni, and
- technology and discovery platforms.

Within the HBS environment, Baker has two main service partners whose work interrelates closely with research data services: (a) the Division of Research and Faculty Development (DRFD), which supports faculty research in areas such as administration, planning, and computing services; and (b) Information Technology (IT), which provides the School's fundamental computing infrastructure.

¹ The publication of the International Association for Social Science Information Services & Technology, which began publishing in 1976 (thus indicating the long history of the field of data services).

² Run by the Digital Curation Centre, which began publishing in 2006, reflecting a recent expansion in this field.

Furthermore, many players within the broader Harvard University community, including the Harvard University Libraries, have been discussing, developing, and promoting an expanded and more robust environment of research data services.

In 2012, building on a history of supporting data for research in a variety of ways, Baker put forth its first formal proposal for a research data management (RDM) program, driven both by user needs and university policies on data security and records retention. The program was intended to establish services and policies to manage research data within three discrete areas:

1. Acquisition and curatorship of data
2. Research methods and tools in creation and manipulation of data
3. Preservation and archiving

The program was designed to be carried out by a centrally coordinated team of Baker staff drawn from a variety of departments involved in data management activities and chosen for their diverse expertise. This team evolved in form and name over time, receiving input and involvement from advisors and stakeholders within Baker and across HBS. In the years since the program's inception, major activities and milestones have included

- environmental scans (of peer institutions and data infrastructure providers);
- a survey of and interviews with HBS stakeholders in order to assess research data management practices and needs;
- gap analyses;
- staff upskilling and learning about RDM;
- hiring staff with designated roles for planning and coordinating research data services;
- broadening the program from one focused on management of data collections to one designed to provide services for customers across the entire research data lifecycle;
- communications to the HBS community about best practices in RDM and supporting library services;
- establishment of the HBS Dataverse (Harvard Dataverse, [2018](#)), a collection within the Harvard Dataverse to publish and disseminate HBS-produced research data; in doing so, the department built a model for staff-supported deposit and recruited appropriate data for publication; and
- creation of an online tool (see Service Highlight) that can identify datasets acquired individually by HBS faculty.

Portfolio of Current Services

At present the Baker Research Data Program (RDP) continues to be a cross-departmental project within the library. The RDP is led by the research data & collections librarian, who coordinates and leads the work of staff across various departments within Baker, in collaboration with the heads of each of those departments, each of whom contributes to the RDP as a component of his or her job. Within Baker, the program's stakeholder group, which includes department heads, managers, and other key involved staff, facilitates program-wide communication and is consulted on the RDP's goals and direction.

Baker currently provides the following services for researchers at each phase of the data life cycle:

- Plan: Staff provide advice on planning research and data management tasks, as well as help to assess the data needs for projects.

- **Collect/generate:** As academic libraries have done historically, staff enable users to find data for their research, by providing both discovery tools (e.g., library search systems and web guides) and individualized guidance from reference staff. Baker staff collect data for research in many forms as part of the Library's collection development. In addition, Baker provides a service, rare among academic libraries, to facilitate and negotiate the purchase of data using faculty research budgets. Furthermore, faculty may enlist Baker staff to search for and pull data for them for a project.
- **Clean/analyze/visualize:** A team of Baker staff with expertise in statistics and data handling support researchers in numerous ways to prepare data for the analysis phase of research. They organize the data (through tasks such as cleaning and merging), do preliminary analyses (such as descriptive statistics and data visualization), and advise on further statistical methodology that the researchers may employ.
- **Publish, share, and archive:** Baker staff will publish HBS-produced research data in the HBS Dataverse. If the HBS Dataverse doesn't meet the faculty member's needs, staff can advise researchers on outside data repository options. Furthermore, Baker Special Collections, in their overall role of collecting and archiving the record of HBS faculty research, archives key research data of the School for long-term preservation and use. Finally, a task of increasing importance is to promote—and track the use of—HBS-published data.

These Baker services fit within a wider context that includes support for research data management from other departments within HBS (such as DRFD and IT) and across Harvard throughout the data life cycle, in areas such as tracking research outputs, primary data collection, research computing, and short- or long-term data storage. Moreover, many departments and stakeholders across HBS contribute to the development of new library products and services, as seen in the next example.

Service Development Highlight: Research Datasets Tool

Over the past 5 years, there has been a growing desire among HBS faculty for greater transparency regarding which research datasets were “owned” or licensed by the school. One faculty member, representing a view shared by many of his colleagues, lamented:

I do not understand why the school buys proprietary datasets for faculty... and there's no central repository of the data or even of the codebooks, or the descriptions of what any of us own.... You might abandon a project because you think the data doesn't exist, but actually someone down the hall bought it two years ago and it exists at HBS and you don't know that.

To help address this need, Baker launched a project in January 2017 to develop a secure, online catalog of datasets which would allow faculty, and their doctoral students, to independently discover previously “hidden” datasets acquired by individual faculty but not otherwise identifiable through standard library discovery tools.

The idea of creating a database referencing datasets was not unprecedented at HBS. Other faculty had suggested the tool in the past, but several obstacles impeded the effort. Until recently, HBS lacked a secure content management system (CMS) with the ability to limit access to the tool to faculty and doctoral students, and to allow staff to tag the resources according to faculty's research needs. In addition, faculty were sometimes reluctant to share information about their datasets thinking exclusive

access might serve as a competitive advantage. Gradually, however, open data initiatives in the social sciences, requirements by scholarly publishers for greater transparency around research data, and the promotion of broader collaboration on research projects by HBS's own Research Directors lay the groundwork for the development of the Research Datasets Tool.

The goals of the Research Datasets Tool were clearly articulated and embraced by key stakeholders at HBS, which included groups within Baker Library, the HBS Research Directors, faculty, and the DRFD. The goals were to

- aid faculty in the discovery of previously "hidden" research data;
- maximize reuse of datasets purchased with HBS funds;
- help identify potential cost-share partners for data purchases;
- enhance research dataset information with vendor notes, licensing restrictions, codebooks, data samples, programming notes, etc.; and
- leverage Baker's new CMS platform to develop a stable and easy-to-maintain research tool.

To develop the tool, Baker Library would use its new Research Innovation Framework—a cross-organizational effort to develop a more structured and responsive approach to selecting, designing, developing, and maintaining library information products and services for priority customers (Dolan, Hemment, & Oliver, [2017](#)).

Utilizing a "design-thinking" approach, a small cross-functional team from Baker, joined by a colleague from the DRFD, kicked off the Research Datasets Tool project in January 2017 with a series of discussions with faculty to better understand their unique needs and challenges around research data usage and discovery. From these meetings, not only did key functional and technical requirements emerge for the tool itself, but valuable perspectives regarding library services that could complement the tool were also identified. Those initial discussions with faculty helped the team to shape the next wave of the Research Data Program and initiate a more proactive outreach to Baker faculty and students.

One of the key functional requirements of the Research Datasets Tool that was repeatedly requested by faculty was the need for specialized filters that would allow them to explore the datasets from a variety of angles—by topic, geographic area, industry, unit of analysis, and date. Other functional requirements for the tool included:

- bookmarking;
- one-click data download (when possible);
- links to data documentation, codebooks, data samples, etc.;
- links to related research;
- identification of primary contact; and
- identification of acquiring researcher.

The Research Innovation Framework process enabled the team to bring together expertise from across Baker Library and to work effectively with external stakeholders. Baker's internal product team included a project manager, product designer, metadata and taxonomy specialists, faculty research staff, and web developers. Based on the requirements identified by HBS faculty and research directors, Baker staff built a functional tool that incorporated a number of key features, which can be seen in the following screenshots.

Figure 1. A screenshot of the Research Databases Tool illustrating some of the key features on the main results page

Raising awareness among faculty about the Research Datasets Tool is an ongoing challenge. When it was first launched nine months ago, an email announcement went out to all faculty promoting the resource. Since then, when new datasets have been added to the tool—a key responsibility of Baker’s Research Data Program staff—they have been showcased in a weekly email newsletter distributed to all HBS faculty. Baker staff also highlight the Research Datasets Tool when meeting with new faculty and during research consultations. Nonetheless, Baker staff regularly encounter faculty who admit to never having heard of or seen the tool. To increase awareness, Baker will shortly commence a promotional campaign with a special focus on academic departments, such as the Finance Department, that have a data-intensive research focus. The Baker team hopes to increase understanding of both the tool itself and how it fits within the broader research data environment.

If effective, this campaign would increase not only use of the tool but also requests coming into the library for access to the data. In order to properly service such increasing demand, Baker’s Research Data Program staff will review the library’s data request process to seek opportunities for greater efficiency and standardization. For example, servicing requests for access to datasets listed in the tool to date often has required a single, designated staff member to review several pieces of documentation, including, in

some cases, a licensing agreement, in order to determine if and how the requestor can have access. Greater standardization and machine-readability in staff-side documentation of data licenses could ease the final step of providing access.

Figure 2. A screenshot of the Research Datasets Tool illustrating some of the key features on the individual datasets page

ABOUT THIS DATASET

Corporate environmental and social responsibility data on approx. 2,000 companies worldwide responding to an information request sent by CDP on behalf of its signatory investors. Datasets include Climate Change Data, Supply Chain Data, Water Data, Forests Data, and CDP Scores Data.

DOCUMENTATION

[CDP Climate Change Survey Questions 2015](#)

DOWNLOAD DATA HERE

LICENSE INFORMATION

Users:	Harvard University
Expires:	No end date

COVERAGE

2003 - 2015

DATA INTERVAL

Annual

FILE TYPE

Excel

UNIT OF ANALYSIS

Organization

CONTACTS

BAKER RESEARCH SERVICES

To request access or provide feedback, contact:

ALEX CARACUZZO
acaracuzzo@hbs.edu
617.495.2492

ACQUIRING RESEARCHER

GEORGE SERAHEIM; JODY GREWAL

RELATED RESEARCH

Ioannou, I., Li, S., and Seraheim, G. "The Effect of Target Difficulty on Target Completion: The Case of Reducing Carbon Emissions." *The Accounting Review* (2016). (PDF, 751.4 KB)

Over the next year, the Research Data Program staff will also assess how well the Research Datasets Tool is meeting its original objectives. Is it, in fact, helping faculty to discover previously "hidden" research data at the school, reuse datasets purchased with HBS funds, and identify cost-share partners for data purchases? How have HBS faculty and doctoral students benefitted and where has the tool fallen short?

In many ways the Research Datasets Tool has already been a resounding success. It has confirmed the efficacy of Baker's Research Innovation Framework approach for new product development, helped the library forge new strategic partnerships with other HBS administrative groups, repurposed existing management information about data acquisitions to create a discovery tool, illustrated the capabilities of

Baker's new CMS and discovery platform, and established the library as the primary contact for services related to licensed research datasets.

Areas of Further Considerations for the Research Data Program

In reflecting upon the past and planning for the future, Baker has identified some key issues that merit further consideration as staff continue to develop the Research Data Program.

To date, Baker has provided valuable services alongside its HBS partners, yet there are many more opportunities to further integrate services across the school in the future, such as at transitional points within the data lifecycle. Furthermore, HBS operates within a distributed university environment, with certain systems and policies set and run at the local (i.e., school) level, and others run centrally at Harvard University. This environment provides the opportunity to both experiment with and roll out tools and services at the local level (exemplified by the Research Datasets Tool), while also seeking ways to partner across Harvard with others looking to foster the university's data management support services.

Data sharing also comes with some particular challenges for business researchers, who collect both primary and secondary data. Some disciplines contend with privacy issues when collecting and sharing data from human subjects, while others utilize data from proprietary sources: business researchers must cope with the perfect storm of both. Common research methods include collecting data directly on persons and/or companies, and companies are often publicly profiled and named, which can complicate the anonymization of participants. Furthermore, many secondary business data sources have very restrictive licenses which may prevent downstream data sharing. These issues require further exploration so Baker can enable data sharing in the most appropriate way for HBS researchers. Solutions may move beyond the binary question of *whether* to share, to establish and articulate various options for long-term management and appropriate access to data. Options may involve a variety of data repositories, various levels of detail in data files, tiered access methods, embargoes, and more.

Data discovery continues to be a challenge for HBS researchers. While the Research Datasets Tool made great headway in increasing transparency, there continues to be a highly distributed discovery environment for research data. Information on additional data available to HBS researchers is contained within Baker and Harvard Library discovery systems, within subscription databases, at various data repositories, and on the vast world wide web. In the absence of a single search system for data, librarians must do their best to point researchers to the most relevant data sources.

Furthermore, Baker is challenged by the task of keeping track of the data published by HBS researchers. While the library is aware of the faculty datasets deposited in the HBS Dataverse, faculty may (and surely do) disseminate their research data via other repositories, such as the ICPSR³, journal data sharing sites, and more. Baker staff have an opportunity to improve tracking of the publication, use, and impact of faculty-produced data wherever it may be published. In this arena, the library can build upon international efforts to track research data as a type of research output.

³ The Inter-university Consortium for Political and Social Research (ICPSR) is the largest and primary archive for research data in the social sciences in the United States. <https://www.icpsr.umich.edu/icpsrweb/>

As Baker grows the RDP, the overarching goal is to provide faculty and other researchers with the most appropriately tailored set of research data services feasible. The needs of researchers will vary based upon the phase of the research life cycle in which they are working and the nature of their data and research projects. Thus, some services will be relevant to faculty at a particular moment, whereas others may be more suitable at a later point or not at all. Yet the organization does not have the resources to provide entirely bespoke services and requires standardization in order to have manageable, scalable processes. Therefore, after new services are explored and piloted, Baker must run them in a standardized way in the form of service modules. Such modules can be matched to each researcher according to his or her needs at a given moment. This model aligns with the concept of mass customization, whereby a researcher can receive the most suitable set of services, which are underpinned by efficient and standardized processes behind the scenes.

Conclusion

Baker staff are currently establishing strategic directions for the next iteration of the Research Data Program. Members of the stakeholders group recently completed an analysis of opportunities, threats, strengths and weaknesses, which generated ideas for the RDP's goals and projects for the coming year. Ideas will be shared and refined with feedback from various stakeholders. The resulting list of goals will evolve over time as the Research Data Program continues to develop.

In the coming years, Baker will build on its existing set of service modules to investigate new opportunities. The library will refine and improve existing tools, workflows, and services to maximize impact. The research data services staff will work to better interconnect the Research Datasets Tool with other library discovery services and improve management of its metadata. The team will consider new areas of service, including integration across HBS, Harvard, and beyond. Baker staff will explore, with HBS research computing services staff, ways to improve data management and enable seamless transitions through the phases of the research life cycle, such as the handoff between active-phase and long-term data storage. Baker staff will collaborate with Harvard Libraries to explore new workflows for depositing data into Dataverse. And finally, Baker staff must examine the library's various internal information systems and determine how to optimize data exchange across Baker, HBS, and Harvard.

The use of data in business is not a new arena, nor are library services to support that work. Yet services must continually adapt to best serve the ever-changing environment at HBS, Harvard, and in the world, as well as evolving researcher needs. As Baker launches the next phase of the Research Data Program, the team aims to enable effective use of data by its researchers in service of the School's mission to educate leaders who make a difference in the world.

References

Dolan M., Hemment M., and Oliver S. (2017). A Framework for Sustaining Innovation at Baker Library, Harvard Business School. *New Review of Academic Librarianship*, 23(2-3), 275-292.

Harvard Dataverse. (2018). Retrieved from <https://dataverse.harvard.edu/dataverse/hbs>

Hswe, P., et al. (2011). Responding to the call to curate: Digital curation in practice at Penn State University Libraries. *International Journal of Digital Curation*, 6(2), 195-208. <https://doi.org/10.2218/ijdc.v6i2.196>

Ribeiro, C., & Fernandes, M.E.M. (2011). Data curation at U.Porto: Identifying current practices across disciplinary domains. *IASSIST Quarterly*, 35(4). <https://doi.org/10.29173/iq893>

Rice, R., Ekmekcioglu, C., Haywood, J., Jones, S., Lewis, S., Macdonald, S., & Weir, T. (2013). Implementing the research data management policy: University of Edinburgh roadmap. *International Journal of Digital Curation*, 8(2), 194-204. <https://doi.org/10.2218/ijdc.v8i2.283>

Rice, R., & Haywood, J. (2011). Research data management initiatives at University of Edinburgh. *International Journal of Digital Curation*, 6(2), 232-244. <https://doi.org/10.2218/ijdc.v6i2.199>

Rumsey, S., & Jefferies, N. (2013). Challenges in building an institutional research data catalogue. *International Journal of Digital Curation*, 8(2), 205-214. <https://doi.org/10.2218/ijdc.v8i2.284>

Schoots, F., Sesink, L., Verhaar, P., & Frederiks, F. (2017). Implementing a research data policy at Leiden University. *International Journal of Digital Curation*, 12(2), 256-265. <https://doi.org/10.2218/ijdc.v12i2.575>

Steinhart, G. (2009). DataStaR: An institutional approach to research data curation. *IASSIST Quarterly*, 31(3). DOI <https://doi.org/10.29173/iq187>

Walters, T.O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *International Journal of Digital Curation*, 4(3), 83-92. <https://doi.org/10.2218/ijdc.v4i3.116>

Wilson, J.A.J., & Jeffreys, P. (2013). Towards a unified university infrastructure: The data management roll-out at the University of Oxford. *International Journal of Digital Curation*, 8(2), 235-246. <https://doi.org/10.2218/ijdc.v8i2.287>

Wilson, J.A.J., Martinez-Urbe, L., Fraser, M.A., & Jeffreys, P. (2011). An institutional approach to developing research data management infrastructure. *International Journal of Digital Curation*, 6(2), 274-287. <https://doi.org/10.2218/ijdc.v6i2.203>