



Global Information Industry Center

at the School of International Relations and Pacific Studies

University of California, San Diego

Chemistry and Chemical Engineering at MIT The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, John Norris Maguire Professor of Information Technology, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering

Mackenzie Smith, Associate Director of Technology, MIT Libraries

Kate Clopeck, Masters of Science, Technology and Policy Program, MIT

June 2009

Abstract:

This case study details how scientists at MIT’s Department of Chemistry Instrumentation Facility (DCIF) create, use and store data at the DCIF. The majority of data created at the DCIF is produced by seven Nuclear Magnetic Resonance (NMR) spectrometers. These are large, highly specialized devices that allow scientists to study the physical, biological and chemical properties of matter. DCIF spectrometers generate a maximum of 3.3 gigabytes of spectral data per week, or approximately 165 gigabytes of data per year. The amount of data generated by the NMR spectrometers is growing as their usage increases. The case study concludes by explaining the DCIF’s data retention policies and the sharing and reuse of data with other university laboratories. Other papers examine other labs at MIT.

- Executive Briefing:** a summary of one or more research projects with preliminary findings for a non-academic audience.
- Research Report:** a completed report drawing on one of more research projects that presents study data, findings and management implications.
- Case Study:** an in-depth description of a firm’s approach to an information management issue.
- Research Article:** an academic research paper with sections on hypotheses tested, methods and data, analysis, findings and references.

Contents

Background	3
Data Generation	3
Metadata	6
Data Retention.....	6
Data Sharing and Reuse	6
Key Trends and Indicators for Data Growth	7
About the HMI? Program	8
Acknowledgements	8

Chemistry and Chemical Engineering at MIT

The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, MacKenzie Smith, and Kate Clopeck, MIT

June 2009

Background

The Department of Chemistry at the Massachusetts Institute of Technology has over 30 faculty members who teach and conduct research on a variety of subjects including biological chemistry, inorganic chemistry, organic chemistry, physical chemistry, environmental chemistry, materials chemistry and nanoscience. One scientist interviewed, for example, is currently studying quantum chemistry in an effort to develop new methods to make reliable predictions about chemical phenomena. Currently, his lab is focused on physical chemistry topics such as electron transfer, electron dynamics, electron spins, and molecular magnetism.

Many faculty members of this department conduct experiments at the Department of Chemistry Instrumentation Facility (DCIF). This NSF-funded facility's function is to maintain state-of-the-art major analytical instruments in order to support the ongoing research programs within the MIT Chemistry Department. Currently, four permanent staff members provide instrument training, maintenance, repair, and applications assistance to well over four hundred users. The lab houses seven Nuclear Magnetic Resonance (NMR) spectrometers, one Electronic Paramagnetic Resonance (EPR) spectrometer, one high-resolution Fourier Transform mass spectrometer, a Gas Chromatograph mass spectrometer, a polarimeter, a Bruker Omnix MALDI-TOF, and a Fourier Transform Infrared (FT-IR) spectrometer.

In addition to the Chemistry Department, MIT has

a separate department of Chemical Engineering. Chemical Engineers at MIT conduct research in areas of chemistry, biology, and physics, and have made significant contributions to the fields of medicine, biotechnology, microelectronics, advanced materials, energy, consumer products, manufacturing, and environmental solutions. For example, one scientist in this department conducts research in areas of metabolic engineering, biochemical engineering, bioprocess engineering, and synthetic biology to harness the synthetic power of biology to build “microbial chemical factories.” Her current efforts are focused on the development of tools and methodologies for novel biosynthetic pathway design and the investigation of gene dosage effects on the physiology and productivity of engineered microbes.

For this case study, three scientists were interviewed, including two from the Department of Chemistry and one from Chemical Engineering.

Data Generation

As mentioned earlier, scientists in the Departments of Chemistry and Chemical Engineering conduct research in various subject areas from quantum chemistry to biotechnology. The amount of data generated by each scientist varies based on the goals of the specific research project and the instruments used to generate data. One way to estimate the total amount of data generated by scientists in these two departments is by the data produced at the Department of Chemistry's Instrumentation Facility (DCIF).

The DCIF is open to faculty members and research groups at MIT, as well as members of other academic institutions and of industry in the area. Over the course of a year, about 60 research groups (over 400 users) actively use the DCIF. Industrial customers use about 12% of the total “instrument use time”, while the remaining 88% is used by groups from academic institutions (MIT research groups account for 84% of “instrument use time” by academic institutions).

The majority of the data generated at the DCIF is produced by the Nuclear Magnetic Resonance (NMR) spectrometers. NMR is a phenomenon that occurs when the nuclei of certain atoms are immersed in a static magnetic field and exposed to a second oscillating magnetic field. Not all nuclei experience this phenomenon – it depends on whether the protons in the nucleus possess a property called spin. The spin of a proton is like a magnetic moment vector, which causes the proton to behave like a magnet with a north and south pole. When the proton is placed in an external magnetic field, the spin vector of the particle aligns itself with the external field, just like a magnet would¹.

Spectroscopy is the study of the interaction of electromagnetic radiation with matter. Nuclear magnetic resonance spectroscopy is the use of the NMR phenomenon to study physical, chemical, and biological properties of matter. NMR spectroscopy is routinely used by chemists to study chemical structures using simple one-dimensional techniques. Other NMR techniques include: two-dimensional techniques to determine the structure of more complicated molecules, time domain techniques to probe molecular dynamics in solutions, and solid state NMR spectroscopy to determine the molecular structure of solids.

Figure 1 is a schematic representation of the major systems of a NMR spectrometer. At the top of this diagram is the NMR spectrometer's super conduction magnet. This magnet is one of the most expensive components of the nuclear magnetic resonance spectrometer system and produces the static magnetic field necessary for all NMR experiments. The shim coils (which are located immediately within the bore of the magnet) are for homogenizing the magnetic field produced by the super conduction magnet. Within the shim coil is the probe, which contains RF coils. The sample is positioned within the RF coil of the probe. These RF coils serve two purposes: 1.) To produce the second, oscillating magnetic field, which is necessary to rotate the spins of the sample during NMR experiments; and 2.) To detect the signal from

¹ Hornak, Joseph P. The Basics of NMR. The Rochester Institute of Technology (2002).

the spins within the sample.

As shown in Figure 1, the instrument's computer controls all components of the spectrometer. The operator of the spectrometer gives input (i.e. RF frequency, the width and shape of the RF electromagnetic pulses for the oscillating magnetic field) to the computer through a console terminal with a mouse and keyboard. Some spectrometers also have a separate small interface for carrying out some of the more routine procedures on the spectrometer. A pulse sequence is selected and customized from the console terminal. The operator can see spectra on a video display located on the console and can make hard copies of spectra using a printer.

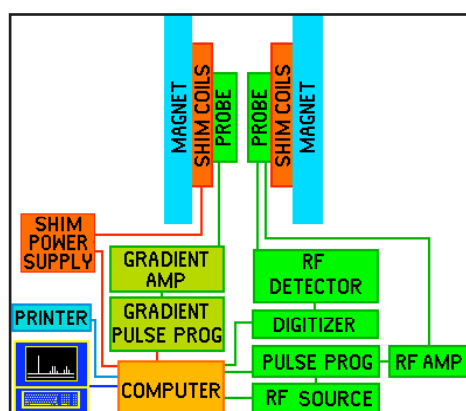


Figure 1: Schematic representation of the major systems of a nuclear magnetic resonance spectrometer¹

NMR spectrometers produce spectral data. Figure 2 illustrates an example of a low resolution NMR spectrum. The number of peaks in the spectrum is equal to the number of different environments the hydrogen atoms are in. The ratio of the areas under the peaks is the ratio of the number of hydrogen atoms in each of these environments. The amount of splitting indicates the number of hydrogens attached to the carbon atom or atoms "next-door." The number of sub-peaks in a cluster is one more than the number of hydrogens attached to the "next-door" carbon(s). Figure 2 shows the NMR spectrum for C₄H₈O₂. The three peaks indicate that there are three different environments for the hydrogens. The hydrogens in those three environments are in the ratio

2:3:3. Since there are 8 hydrogens altogether, this ratio represents a CH₂ group and two CH₃ groups. The CH₂ group at about 4.1 ppm is a quartet, which means that it is “next-door” to a carbon with three hydrogens attached - a CH₃ group. The CH₃ group at about 1.3 ppm is a triplet. Therefore, this group must be “next-door” to a CH₂ group. The CH₃ group at about 2.0 ppm is a singlet. That means that the carbon “next-door” is not attached to any hydrogen atoms².

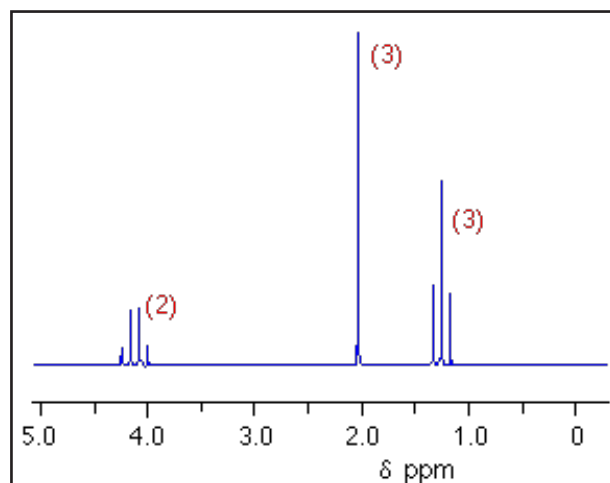


Figure 2: Low resolution NMR spectrum for C₄H₈O₂.

In total the 7 NMR spectrometers instruments at the DCIF generate a maximum of about 3.3 gigabytes of data per week or approximately 165 GB of raw data per year (see table 1). The biggest data generator is the Bruker AVANCE 400 MHz NMR spectrometer with Spectro Spin superconducting magnet. This instrument generates approximately 677-843 MB per week. Although the other instruments at the facility do generate data, it is insignificant when compared to the amount produced by the NMR spectrometers (about 1/10th of the data).

Table 1: Data generated by the NMR spectrometers at MIT’s DCIF

NMR Spectrometer	Data Generated (MB/week)
VARIAN Mercury 300	217-251
Bruker AVANCE-400	582-751
Bruker AVANCE-401	677-843
VARIAN Inova-500	259-412
VARIAN Inova-501	103-176
VARIAN Inova-502	114-216
Bruker AVANCE-600	135-661

The amount of data generated at the DCIF has increased over time. This increase is due to an increase in the use of the instruments, not a change in the instruments themselves. Therefore, the best way to gauge this increase is by examining the billing statements, which describe the amount of minutes the facility bills each month. In 2003, the facility billed an average of 99.5 kilominutes per month and a total of 1194 kilominutes for the year. In 2008, the monthly average increased to 115 kilominutes and the total billed time increased to 1382 kilominutes. Since there is a limit to the amount of time that the instruments can be used, the data generated at the DCIF has the potential to plateau (assuming that the facility does not add any new instruments). However, based on current trends, the facility does not expect to reach this plateau in the next five years, and predicts the increase in billed time to be similar to the increase over the past five years. This would result in a monthly average of 133 billed kilominutes and an annual total of 1595 billed kilominutes in the year 2014.

Although many faculty members in the Chemistry Department use the DCIF, others may conduct experiments at different facilities, either at MIT or other collaborating institutions. Additionally, there are a number of scientists in this department who generate data from models and simulations and do not conduct experiments at all. These scientists often generate a lot of “intermediate” or “temporary” data while their models run, but only need to save a small amount of this data in the end. For example, the scientist in this department studying quantum chemistry generates approximately 32 gigabytes of temporary data per run but then condenses that

² <http://www.chemguide.co.uk/analysis/nmr/highres.html>

data into about 1 megabyte of what he considers “output data.” The temporary data usually contains more information than the research group is able to analyze or store, so they sift through and keep only the information needed for the given project. For the quantum chemist, the research data is usually the x, y, and z coordinates of an atom in a certain chemical situation, while the temporary data would usually include “everything you would want to know about the atom.” Although he does not generate large amounts of data, the data production has increased over time. Five years ago he stored approximately 100 gigabytes of data, and today he has about 1 terabyte of total data stored. In the future, he predicts that data generation will continue to increase as computers get faster and storage prices decrease.

Metadata

Like the raw experimental or model data, the metadata generated by chemists and chemical engineers at MIT varies based on the research project. For example, the metadata for the chemist mentioned in the previous section is mostly descriptors of the accuracy of the model used. This information is stored in text files that are usually about 1 kilobyte per project. This metadata is often stored in lab notebooks as well.

Data Retention

There are no formal data retention policies for the Chemical Engineering or Chemistry Departments. Each scientist decides how to store, manage and back up his or her own data. One chemical engineer, who generates High Performance Liquid Chromatography (HPLC) and mass spectrometry data, will store all of her data in two locations: on the HPLC instrument, and on her students’ personal computers. Each student is in charge of his or her own data. Lab members use MIT’s central back up service to back up the data on their personal computers, but not the data on the instrument computers. While MIT’s back up service does offer nightly back up, most students are not constantly connected to the MIT network and need to manually back-up their data on

their own schedule. In addition to the data storage on computers, this scientist also keeps hard copies of the data that she personally generates. First she records all of the data in her lab notebook, and then she prints out copies of all of the data generated by an experiment. While she believes in creating paper back ups, not all of the students in her lab practice this technique.

While the Chemical Engineering and Chemistry departments as a whole do not have specific data retention policies, since their disk space is limited the DCIF is trying to institute a five-year data retention policy for the facility. Currently, this facility has approximately 1.5 terabytes of disk space on the main lab computers. If they get close to capacity, then the facility will delete the oldest files. If the new five-year retention policy were implemented, then the facility would automatically delete any data that has been stored on the main lab computers for more than five years.

The DCIF backs up all of the data that is generated in their lab onto DVDs. Each week they copy two weeks worth of data (so each week’s data is eventually copied twice). The DVDs are kept onsite. Based on the data generation estimates discussed in the previous section, this facility backs up about 330 GB of data per year.

Data Sharing and Reuse

While many chemists and chemical engineers from different laboratories do share data there are no widely used national data repositories. For the experiments run by the chemical engineer mentioned in the previous section, the experimental conditions are often more important than the results. She will often call her colleagues at different universities or labs and ask about their experimental conditions.

Another scientist in the Department of Chemistry sometimes uses the National Institute of Standards and Technology’s Computational Chemistry Comparison and Benchmark Database. This database is a collection of experimental and ab initio thermochemical properties for a selected set

of molecules. The goals of this data collection are to: 1) provide a benchmark set of molecules for the evaluation of ab initio computational methods; and 2) allow the comparison between different ab initio computational methods for the prediction of thermochemical properties. The thermochemical values included in the CCCBDB are enthalpies of formation, entropies, heat corrections (integrated heat capacity), data needed to compute thermochemical properties (such as geometries, rotational constants, vibrational frequencies, barriers to internal rotation, and electronic energy levels), and additional computed properties (such as atomic charges, electric dipole moments, quadrupole moments, polarizabilities, and HOMO-LUMO gaps)³.

Key Trends and Indicators for Data Growth

Although the different researchers in the Departments of Chemistry and Chemical Engineering study a range of different topics, there are several key trends and indicators for data growth that we have identified:

1. MIT's Department of Chemistry Instrumentation Facility generates approximately 165 GB of data per year, and an additional 330 GB of data backups. The majority of this data is generated by the seven Nuclear Magnetic Resonance (MR) spectrometers. Of these seven NMR spectrometers, the biggest data generator is the Bruker AVANCE 400 MHz NMR spectrometer with Spectro Spin superconducting magnet. This instrument generates approximately 677-843 MB per week.
2. 12% of the total "instrument use time" at the DCIF is from industrial customers. The remaining 88% is used by groups from academic institutions (MIT research groups account for 84% of "instrument use time" in this category).
3. In addition to the scientists using the DCIF, there are also many chemist and chemical engineers that generate data with models instead of instruments. While these scientists can generate

large amount of data while the model is running, most of this data is the temporary outputs of calculations and is either deleted or significantly condensed before the model is finished running.

4. Data retention and back-up policies vary depending on the preference of the specific researcher. Some scientists will delete their data after publishing a paper, while others keep everything until their storage capacity is reached (and then delete the oldest files). Others will keep their data forever and buy more storage if they reach their current storage capacity.

3 <http://cccbdb.nist.gov/>

About the HMI? Program

The How Much Information? (HMI?) research program is a multi-discipline, multi-university project, formed to investigate the nature of data and information generated and used by individuals and enterprises. The program is sponsored by seven companies, including AT&T, Cisco, IBM, Intel, LSI, Oracle, and Seagate, and involves multiple research universities. The Principal Investigator is Prof. Roger Bohn and the Research Director is Dr. James Short, at UC San Diego's Global Information Industry Center (<http://giic.ucsd.edu>). Founded in 1960, the University of California, San Diego is one of the nation's most accomplished research universities, widely acknowledged for its local impact, national influence and global reach.

Acknowledgements

This case study is the product of industry and university collaboration in applied research. We are grateful for the support of our industry partners, sponsor liaisons, university research partners, and administrative staff at the University of California, San Diego.

Financial support for HMI? research and the Global Information Industry Center is gratefully acknowledged. Our sponsors are:

AT&T

Cisco Systems

IBM

Intel

LSI

Oracle

Seagate Technology

Additional support was provided by the Alfred P. Sloan Foundation of New York.

Questions about this research may be addressed to the Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego:

Roger Bohn, Principal Investigator, rbohn@ucsd.edu

Jim Short, Research Director, jshort@ucsd.edu

Pepper Lane, Program Coordinator, pelane@ucsd.edu