**Going to the cloud vs. doing it in-house**

by Pamela Carson (Web Services Librarian)
Kathleen Botter (Systems Librarian)
Stephen Krujelskis (Systems Administrator/Analyst)

Concordia University Libraries, Montreal
Submitted to Computers in Libraries on May 3, 2013

**Introduction**

Cloud computing, broadly defined as a third-party information technology services delivered via the internet, has the potential to revolutionize library IT departments. In fact, your library might already be using aspects of the “cloud” such as SaaS (software-as-a-service) or IaaS (infrastructure-as-a-service), or less commonly PaaS (platform-as-a-service). The Primary Research Group (2011) published a report based on a survey of over 70 libraries worldwide which showed that over 60 per cent of libraries are already using free or paid SaaS while four per cent used IaaS. The popularity of SaaS can perhaps be linked with the common mass-market use of highly usable cloud-based services such as free e-mail like Gmail and cloud storage such as Dropbox. Luo (2013) found that over 50 per cent of reference librarians in the U.S. used cloud-based video services, information collection services, and calendar services. EDUCAUSE found that practically all undergraduate students use cloud-based technologies (Smith & Caruso, 2010). Cloud computing is here and cloud-based solutions for library IT are emerging, but there are important factors to weigh when considering whether the cloud is right for your library.

Concerning IaaS, several factors are important when thinking of moving to the cloud. IT support is different with cloud-based services. Current in-house IT staff will have added duties like negotiating contracts and understanding more legal implications. Backup and disaster recovery practices differ for cloud-based data, and new problems emerge with regard to contingency planning because of the introduction of third parties. Networks and bandwidth and the amounts of data need to be factored in as well. With SaaS, other support- and reliability-related aspects come into play.

Libraries are undoubtedly concerned with patrons’ privacy. Questions of jurisdiction are raised when storing data in the cloud such as vulnerability to the PATRIOT Act, and warrants or subpoenas in other jurisdictions. There is also the importance of negotiating beneficial contracts, the possibility of using HIPAA-like standards for de-identification of user data, as well as concerns of data leakage and linkage.

Integrated library systems (ILSs), a core capacity of libraries, can and have been moved to the cloud, but special security and privacy aspects are important to consider.

**IaaS vs. in-house infrastructure**

There are several factors to evaluate when deciding between in-house infrastructure and cloud-based IaaS in libraries. For IaaS, there are issues of pricing and support, backup and discovery recovery, contingency planning for outages, networks and bandwidth usage, and data storage. For software as a service (SaaS) commonly used in libraries there are questions of support as well. This section discusses the factors to be considered for both in-house and cloud-based cases.

**Pricing and support**

IaaS providers benefit from economies of scale; because they order hundreds or thousands of servers per year, they have far more negotiating power than a library that may buy a handful of servers every few years. These (reduced) costs are spread over multiple customers, keeping prices and barriers to entry for small businesses low. Support is also more efficient for IaaS providers. They are likely to have parts in stock and standard hardware configurations allowing them the ability to address problems very quickly. However, support does not tend to be a major point in deciding between cloud-based and in-house. For example, in the event of a hardware failure affecting library servers at our university, we call the manufacturer and the parts are guaranteed to be delivered within four hours. In a large city like ours (Montreal), frequently replaced parts are likely kept in stock at the courier. Our experience has been that parts are delivered within two hours rather than four.

Libraries will always need systems analysts, with or without the cloud. Some aspects of the sysadmin’s job may be made easier with the cloud – the ability, for example, to choose from a variety of pre-configured system images with common application bundles (e.g. LAMP stack). The downside of these images is that they are often not supported by the cloud service provider. Systems librarians’ roles also change with the adoption of cloud computing, with a shift in focus from technical activities to understanding service agreements.

**Backup and disaster recovery**

Some cloud services guarantee that your data will be replicated to different geographic locations. This would be something that should be verified in the contract. Fully understanding exactly what is being backed up (application, data or both?) and where it is being kept (one location or multiple locations?) are very important.  However, backing up data to the cloud could be a challenging proposition. Even attempting to do backup and disaster recovery within your institution by cooperating with other departments can be hard if you have to argue the case for physical space in data centers and support. The ability to achieve redundancy within your organization may come down to having good rapport with staff in other departments.

If you encounter a major problem in your library’s datacenter it is more likely that things would be up and running sooner, because you would have dedicated people on-site with detailed knowledge of your systems. However, if there is a major problem in a cloud datacenter, you are one client among thousands, and may not be first in line. Having specialists in-house to deal with major problems costs money up front, but may end up saving money long term. These are the same drawbacks encountered by organizations that have information hosted by a third party. In-house expertise is still needed because since the user is free to do what they want they are equally responsible for issues that may come up (Galvin & Sun, 2012).

**Contingency planning for outages**

Relying on cloud services for core operations introduces several more possible weak points in delivery. If core applications and data are off-site, organizations are much more vulnerable to outages, because they are relying on third-party service providers to deliver these services. If applications are on the organization’s own network, users can still work even if the internet connection is down. Many organizations only have one internet link, and would be temporarily out of business if that connection were to fail. Moving to the cloud still imposes requirements of redundancy. To ensure access to their hosted core applications, organizations should consider upgrading their Internet connections in order to have multiple paths. Galvin and Sun (2012) also stated that putting key library applications, such as an ILS, in the cloud requires a full contingency plan due to concern over the reliability of the connection between an IaaS provider and the campus.

**Networks and bandwidth usage**

Universities, colleges, research institutes, hospitals and government laboratories are fortunate to have dedicated, high-speed, high-capacity networks in both the U.S. and Canada, which offer lots of bandwidth for research and innovation. For example, the equivalent to Internet2 in Canada is CANARIE: Canada’s Advanced Research and Innovation Network. Concordia connects to CANARIE via RISQ, an ultra-high speed network in the province of Quebec with over 6000 kilometers of fiber-optic cables (RISQ, 2011). Also, our two campuses are directly linked by fiber, forming our own private network off of the Internet. Due to these features, commercial IaaS is not attractive for us. In fact, it would cost more for bandwidth if we were to switch to the cloud. One model that would be interesting would be a consortial private cloud linked to CANARIE, which would provide services for those on the network. IaaS might also be interesting for new projects at the library since cloud computing can offer quick and flexible solutions. For example, Omeka (an open source web-publishing platform for mainly archives) would be a good candidate to run on the cloud (Galvin & Sun, 2012).

**Data storage**

Much of the data Concordia Libraries handles now is text-based. As of early 2013, the totality of our digital operations (including ILS, websites, course reserves, research repository, streaming media service and more) occupied only about 5 terabytes.  Even if we were to venture into storing data sets and/or audiovisual content, there is no reason to separate this content from the CANARIE and RISQ networks by putting it in the cloud with Amazon Web Services or a similar cloud-based service provider. Similarly, since we already have well-equipped data centers in our institution it makes less sense to start moving to the cloud. Galvin and Sun (2012), also writing about the cloud in the context of academic libraries, propose that “the ideal scenario might be IaaS delivered through central IT to departments on an academic campus” (p. 418). The cloud might make sense if a library was starting from scratch and did not want to invest (or did not have) capital to build data centers, like the Rebuilding Higher Education in Afghanistan project led by the University of Arizona libraries where Koha (an ILS) was migrated and hosted in the cloud (Han, 2010).

**Free SaaS concerns**

Support for cloud-based software-as-a-service can be tricky. Some SaaS would certainly be managed through service level agreements (SLAs), but other SaaS are really meant to be mass market tools with no guarantee of dependability, particularly free SaaS tools. For example, at Concordia we started using Delicious.com in 2008 to create feeds for bookmarks on our website. This was a quick and easy solution for librarians editing subject guides because it allowed them to skip any code editing and simply add content to be displayed on feeds set up by the Web Team on their HTML-based research guides. However, in September 2011, a few months after Delicious was sold to AVOS Systems, Delicious feeds stopped working completely. We tried contacting Delicious or finding an FAQ to fix the problem, but all we could find were comments online from others experiencing the same issue. This posed a problem because pages with extensive Delicious feeds were needed for a series of information literacy classes. The service failed us and we started looking for alternatives. After another outage in early 2012 and further issues with feed pages being slow to load, we decided to install Semantic Scuttle (http://semanticscuttle.sourceforge.net/) on our servers and use this as an in-house alternative to Delicious. Other librarians have also had issues with Delicious, complaining that an updated interface was enough to get them to switch to Diigo (Luo, 2013, p. 160). In this situation, we moved from cloud to in-house.

**Deciding**

Cloudorado is a cloud computing price comparison engine (Cloudorado, 2013) that helps you calculate the best option for cloud computing service provider based on how much RAM, storage, CPU power you need and which operating system (Linux or Windows) you prefer. It deals with the basic questions, but other factors such as the ones raised here in this article – support and details about backup and disaster recovery – would have to be investigated separately. Each library is different in terms of in-house expertise, current support agreements with vendors, network and bandwidth situations, as well as storage needs. Each of these items needs to be weighed when deciding between in-house and IaaS. Perhaps a brand new public library with a reliable and redundant Internet connection, no data center, and little in-house expertise, would do well with IaaS, but an established academic library using the research network rather than the Internet, with high-quality institutional IT infrastructure and talented in-house experts does not have much of a reason to move to the cloud, particularly when questions of privacy, security, and reliability are raised.

**Privacy and the cloud**

To quote Donald Rumsfeld (2002), there are known unknowns and unknown unknowns. When using a third party to deliver services or collect information for your library, your library’s grasp on privacy gets just a little bit slipperier. Google CEO Eric Schmidt argued that privacy is a non sequitur in the Internet age (Fried, 2009), but libraries have long supported the protection of users’ privacy and confidentiality. There are several items to think about when considering moving sensitive information to the cloud or integrating cloud-based services with existing in-house services.

**USA PATRIOT Act**

The introduction of the PATRIOT Act had a chilling effect on libraries in the U.S. and beyond. The implications of this act continue to present day. There are fears that cloud service providers based in the U.S. would be compelled to disclose data to the U.S. government under the PATRIOT Act.  In this act, it is stated that “No person shall disclose to any other person (other than those persons necessary to produce the tangible things under this section) that the Federal Bureau of Investigation has sought or obtained tangible things under this section,” meaning that if a U.S.-based cloud service provider was compelled to produce information for an investigation under the PATRIOT Act that it would have to do this and not notify anyone else (USA PATRIOT Act, 2001, Sec. 215). Presumably, this would mean that if a library stores data in the cloud that this data could be accessed by the government without the library’s knowledge.

To get a grasp on how much this happens, there is some indication to be found in transparency reports released by cloud service providers. Google did not release specific numbers, but provided a range of how many National Security Letters (NSLs) received under the PATRIOT Act and a range of how many users and accounts were affected since 2009 (Google, n.d.). For example, in 2012 Google received fewer than 1000 NSLs and between 1000 and 2000 accounts were affected (Google, n.d.). Microsoft and Twitter have also recently published transparency reports for the first time (Microsoft, 2012; Twitter, 2012).

Questions of legal jurisdiction are raised when non-U.S. organizations want to store data in a U.S.-based cloud (Saleh Rauf, 2011). In fact, Canadians have been known to use the PATRIOT Act as an excuse to avoid U.S.-based cloud computing despite the fact that similar anti-terrorism laws exist in Canada under the Canada Anti-Terrorism Act (Kavur, 2010). For example, even if data is stored in Canada, if police need to obtain personal information for an investigation or during an emergency they may not be required to obtain consent to collect it (Office of the Privacy Commissioner of Canada, 2009).

More frequently, user information may be requested by authorities for reasons other than terrorism or espionage covered under the PATRIOT Act. In a recent report by the Electronic Frontier Foundation, several cloud service providers were applauded for protecting users’ privacy (Dropbox, Google, and Microsoft), whereas others were left wanting (Amazon, Apple, and Yahoo!). Neither Amazon, Apple nor Yahoo! requires warrants supported by probable cause to access content (though Dropbox, Google, Microsoft, and others do). Some companies also tell users about government data requests giving users a chance to defend themselves before data is handed over (Twitter, Foursquare, SpiderOak, WordPress, and Dropbox) (Cardozo, Cohn, Higgins, Hofmann, & Reitman, 2013). This is of particular concern to libraries because sensitive patron information, including reading history, is a typical part of any ILS and cloud computing poses risks to this information.

**Consent**

A recent study estimated that in order to read all website privacy policies encountered in one year it would take 201 hours annually – equaling $3,534 – per American Internet user (McDonald & Cranor, 2008-2009, p. 565). The authors of that study encouraged organizations to make privacy policies more easily readable and to present privacy-related information at relevant times. Canada’s federal Personal Information Protection and Electronic Documents Act (PIPEDA) is a Canadian act covering data privacy and is similar in some ways to Health Insurance Portability and Accountability Act (HIPAA) but has a wider reach than health-related records. A case study on the application of PIPEDA with regard to moving personal information beyond Canadian borders stated that user consent was *not* required when e-mail accounts were moved from Canadian to third-party American data storage. The original consent granted by the user when signing up for the service was sufficient since the “use” of the data did not change.

Additional consent would be necessary if the purposes for which that information would be used were to change. For example, if the Canadian organization was to outsource the processing of personal information it would be required to provide notice of the change and details of the service-provider arrangements, also highlighting potential impacts on user information confidentiality (Office of the Privacy Commissioner of Canada, 2008).

**The importance of contracts**

When outsourcing data management to a third-party firm, libraries can formalize the measures taken to protect sensitive information by spelling out terms in contracts. However, contracts cannot override the laws of the country in which the information resides, for example a Canadian organization could not use contractual means to counteract American laws (Office of the Privacy Commissioner of Canada, 2008).

One important clause in the contract would be the mandatory disclosure of security breaches. In Canada, organizations are not required to disclose whether data has been breached.  In the U.S., 46 states have enacted separate legislation with regard to disclosure of security breaches of personal information (National Conference of State Legislatures, 2013), though details vary widely by state and no federal legislation exists as of early 2013 (aside from HIPAA, which only pertains to health information). Bill C-475, recently proposed in Canadian parliament, would make it law that security breaches be disclosed and penalties be imposed for compliance failures (Geist, 2013); though currently in Canada any disclosure of data security breaches is voluntary. Requiring third-party service providers to disclose data security breaches in the cloud through contracts, if possible, is a step in the right direction toward protecting the privacy of library patrons.

**De-identification**

For the U.S. healthcare industry, HIPAA (Health Insurance Portability and Accountability Act) outlines what data can be made available for inspection by third parties. Perhaps this is a model that could be adopted in libraries whereby only anonymized data would be available to third parties (Nicholson & Arnott Smith, 2007). It would involve “applying data warehousing techniques to separate operational data from archived data,” and once data is no longer in the operational phase, “fields containing personally identifiable information can be removed” (Nicholson & Arnott Smith, 2007, p. 1205) and potentially moved to cloud-based storage, avoiding privacy concerns. However, ILS may not be evolved enough to support external patron records.

**Data leakage and linkage**

The “leakage” and “linkage” of private information is a risk when using third-party services on your website. “Web leakage” pertains to the disclosure of a web user’s personal information to third-party sites without the user’s consent or knowledge. Frames or inserts on a web page (for example, an advertisement or a map) can collect information such as location, IP address and browser information (Office of the Privacy Commissioner of Canada, 2012a). It is possible that this information can be linked with other data collected to form a user profile, which could lead to anything from annoying targeted marketing to identity theft. This is certainly a concern when considering using third-party SaaS applications. Research by the Office of the Privacy Commissioner of Canada showed that third-party organizations (online marketing, web analytics, load balancing/content delivery, website performance monitoring/management, social networks, marketing, and digital advertising) received various types of personal data via web leakage (email addresses, names, usernames, postal codes, locations, and other data) (Office of the Privacy Commissioner of Canada, 2012b). It is common to use a third party for analytics work. This passes on user information. Search terms used or pages viewed could hypothetically be linked with other browsing information and identify private information about a specific user (Krishnamurthy, Naryshkin, & Wills, 2011).

**Security**

In the cloud or not, our sensitive data can still be hacked. There are advantages and disadvantages to both types of storage. A large cloud service provider, like Amazon or Microsoft, is more likely to have a very talented team of security experts that could not be matched in-house at your institution. However, being a large cloud service provider may make these companies more of a target for hackers. Trying to get details on the security infrastructure, security audits, and training offered to employees at cloud-based service providers may be challenging (Cervone, 2011). Staying out of the cloud does not necessarily offer more protection. Team GhostShell breached the servers of one hundred major universities worldwide in fall 2012 in its “ProjectWestWind” scandal, leaking 120,000 records (Huff, 2012). There are risks inherent in both in-house and cloud data storage that need to be weighed carefully and in context.

**Library systems in the cloud**

So what does all this mean for the library?  In a cloud based environment the systems librarian’s focus changes from less technical aspects to understanding and managing service agreements, and customization of systems.

Proprietary ILS/library service providers such as Innovative Interfaces, Ex Libris, SirsiDynix, and Serials Solutions, all offer a range of services from complete hosting and system management, to software- and support-only solutions.  There are also hosting and support options for open source ILS such as Evergreen and Koha, discovery platforms, and other web apps for libraries for services like chat reference and room booking.

With regard to security, our holdings information and websites are already freely available online via the OPAC and services like WorldCat and Google Scholar. Gaining unauthorized access to this information is not of the highest concern.

What is of concern, however, is patron information including addresses and telephone numbers, reading history (in some cases), and, more rarely, login information.  Hacking of patron records which include login information could expose not only the patron’s record, but also access to the library electronic database collection, not to mention other systems such as web email, or portal.  External patron verification through use of LDAP or SSO solution, which would take the patron login information out of the patron record, is preferable.

**Conclusion**

Regardless of how libraries decide to go -- in-house or cloud-based -- an adaptive IT services evaluation strategy capable of change and responsive to the speed with which technology evolves is increasingly important. Cloud-based services may not be right at the moment, but in five years they could answer important needs. Other interesting arrangements to consider could be institution-wide or consortial private clouds which could allow for efficiencies and build on existing networks and bandwidth allowances.

References

CANARIE (2012). Retrieved from http://www.canarie.ca

Cloudorado (2013). Retrieved from http://www.cloudorado.com

Cardozo, N., Cohn, C., Higgins, P., Hofmann, M., & Reitman, R. (2013, April 30). Who has your back? Which companies help protect your data from the government. *Electronic Frontier Foundation.* Retrieved from https://www.eff.org/sites/default/files/filenode/who-has-your-back-2013-report.pdf

Cervone, H. F. (2011). Cloud computing: Pros and cons. In E. M. Corrado, & H. L. Moulaison (Eds.), *Getting started with cloud computing: A LITA guide* (pp. 29-35). New York, NY: Neal-Schuman Publishers, Inc.

Fried, I. (2009, December 10). Mozilla worker touts Bing over Google, citing privacy. *News.CNET.com.* Retrieved from http://news.cnet.com/8301-13860\_3-10413473-56.html

Galvin, D., & Sun, M. (2012). Avoiding the death zone: Choosing and running a library project in the cloud. *Library Hi Tech, 30*(3), 418-427.

Geist, M. (2013, February 27). NDP MP Charmaine Borg tries to kickstart Canada’s dormant privacy reform. *MichaelGeist.ca.* Retrieved from http://www.michaelgeist.ca/content/view/6794/99999/

Google (n.d.). *Transparency report: User data requests: United States.* Retrieved from http://www.google.com/transparencyreport/userdatarequests/US/

Han, Y. (2010). On the clouds: A new way of computing. *Information Technology and Libraries, 29*(2), 87-92.

Huff, S. (2012, October 2). Hackers ‘Team GhostShell’ leak 120,000 records from 100 major universities. *Betabeat.* Retrieved from http://betabeat.com/2012/10/hackers-team-ghostshell-leak-120000-records-from-100-major-universities-in-project-westwind/

Kavur, J. (2010, July 5). Don’t use the Patriot Act as an excuse. *IT World Canada.* Retrieved from http://www.itworldcanada.com/news/dont-use-the-patriot-act-as-an-excuse/141033

Krishnamurthy, B., Naryshkin, K., and Craig, W. (2011). Privacy leakage vs. protection measures: the growing disconnect. *WISP.* Retrieved from http://www.goodtimesweb.org/documentation/2012/w2sp11.pdf

Luo, L. (2013) Reference librarians’ adoption of cloud computing technologies: An exploratory study. *Internet Reference Services Quarterly, 17*(3/4), 147-166.

McDonald, A. M., & Cranor, L. F. (2008-2009). The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society, 4*(3), 543-568.

Microsoft (2012). *Law enforcement requests report.* Retrieved from http://download.microsoft.com/download/F/3/8/F38AF681-EB3A-4645-A9C4-D4F31B8BA8F2/MSFT\_Reporting\_Data.pdf

National Conference of State Legislatures (2013). 2012 security breach legislation. Retrieved from http://www.ncsl.org/issues-research/telecom/security-breach-legislation-2012.aspx

Nicholson, S., & Arnott Smith, C. (2007). Using lessons from health care to protect the privacy of library users: Guidelines for the de-identification of library data based on HIPAA. *Journal for the American Society for Information Science and Technology, 58*(8), 1198-1206.

Office of the Privacy Commissioner of Canada (2008). *Findings under the Personal Information Protection and Electronic Documents Act (PIPEDA): PIPEDA Case Summary #2008-394: Archived: Outsourcing of canada.com e-mail services to U.S.-based firm raises questions for subscribers.* Retrieved from http://www.priv.gc.ca/cf-dc/2008/394\_20080807\_e.asp

Office of the Privacy Commissioner of Canada (2009). *Your guide to PIPEDA: The Personal Information Protection and Electronic Documents Act.* Retrieved from http://www.priv.gc.ca/information/02\_05\_d\_08\_e.pdf

Office of the Privacy Commissioner of Canada (2012a, September). *Infographic: How does web leakage happen?* Retrieved from http://www.priv.gc.ca/resource/tool-outil/infographic/wl\_info\_201209\_e.asp

Office of the Privacy Commissioner of Canada (2012b, September). *Web leakage research by the Office of the Privacy Commissioner of Canada – Test Results.* Retrieved from http://www.priv.gc.ca/information/research-recherche/2012/wl\_201209\_e.asp

Primary Research Group. (2011). *Survey of library use of cloud computing.* New York, NY: Primary Research Group.

RISQ (2011). Retrieved from http://www.risq.qc.ca

Rumsfeld, D. (2002, June 6). Secretary Rumsfeld press conference at NATO headquarters, Brussels, Belgium. Retrieved from http://www.defense.gov/transcripts/transcript.aspx?transcriptid=3490

Saleh Rauf, D. (2011, November 29). PATRIOT Act clouds picture for tech. *Politico.com.* Retrieved from http://www.politico.com/news/stories/1111/69366.html

Smith, S. D., & Caruso, J. B. (2010). *ECAR study of undergraduate students and information technology.* EDUCAUSE. Retrieved from http://www.educause.edu/library/resources/ecar-study-undergraduate-students-and-information-technology-2010

Twitter (2012). *Twitter transparency report.* Retrieved from https://transparency.twitter.com/

Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT Act) Act of 2001, Pub. L. No. 107 –56. 115 Stat. 288 (2001). Retrieved from http://www.gpo.gov/fdsys/pkg/PLAW-107publ56/pdf/PLAW-107publ56.pdf