MDR, Vol. 41, pp. 97–102, December 2012 • Copyright © by Walter de Gruyter • Berlin • Boston. DOI 10.1515/mir-2012-0014

Access and Curation of Digital Cultural Heritage in the National and University Library of Iceland

Ingibjörg Steinunn Sverrisdóttir, Kristinn Sigurðsson, Örn Hrafnkelsson

Ingibjörg Steinunn Sverrisdóttir (iss@landsbokasafn.is) is National Librarian, Kristinn Sigurðsson (kristinn@landsbokasafn.is) is Head of IT and Örn Hrafnkelsson (orn@landsbokasafn.is) is Head of National collections and digital conversion at the National and University Library, Reykjavik, Iceland

The National Library of Iceland (established 1818) and the Library of the University of Iceland (established 1940) were amalgamated in 1994 as the National and University Library of Iceland (NULI). When the changes occurred, an opportunity to transform operations and services opened. It was done by rationalizing and modernizing the functions of the two libraries in a new building and developing new strategies and priorities (Sigurðsson 2000). Since then the mission of the Library has been to serve the users in a new way. Access has been given to both primary and secondary sources within the collections, sources which sometimes are unique or the only exemplars known.

It was decided to aim for the use of the newest technology available and to initiate the development of a digital library, even though it would be smallin the beginning. This required periods of learning and experiments on behalf of the staff, but knowledge in the field started to accumulate (Hannesdóttir 2005).

Digitization

NULIs digitization efforts began with the project *Antique Maps of Iceland* (Islandskort.is). The goal was to make digital copies of all old maps of Iceland in the possession of the Library and make them available on the Internet.

The main reason for selecting the maps was conservation. Maps are not easy to handle, they can be large and sometimes have to be folded, and they are sensitive to external influences like heat and moisture. The best preservation policy is to digitize them, allowing the user to view and explore the digital copies instead of the originals.

Another reason for selecting the maps was their limited number, making it easier to manage the project. In the first phase, 220 maps made between 1540 and 1900 were digitized. The work started in 1996 and the web was opened in 1997.

The original project was in collaboration with Nordic Digital Library Center (NDLC) and received a grant from the Nordic Council for Scientific Information (Nordinfo). The staff at NULI took care of photographing, historical research, classifying and cataloging. The photographs were sent to NDLC in Norway and converted into digital format (Forn Íslandskort 2000). Short historical descriptions in Icelandic and English were based on the book *Kortasaga Íslands* (A history of cartography of Iceland) by Haraldur Sigurðsson (Sigurðsson 1971-78).

Over the years the project and the website have undergone a total renovation. The maps have been digitized again in NULI, including the backsides of the maps. The new scans produced much higher resolution images, bringing a wealth of details that had been missing from the older images. The metadata has been updated and linked to the national library system, Gegnir.is and the national discovery service, Leitir.is. The scope of the project has also expanded, whereas previously the collection was limited to maps made before 1900, it will now stretch to 1944. The new edition was launched in October 2012.

The next digitization project, *Sagnanet*, was on a much bigger scale. It was carried out in cooperation with Cornell University and the Árni Magnússon Institute for Icelandic Studies and sponsored by the Andrew W. Mellon Foundation, the Icelandic government and several other Icelandic partners. Digitization began in 1997 and the website was opened in 2001. The aim was to give access to digital images of about 220 thousand manuscript pages and 151 thousand printed pages (Hallgrímsson 2006a).

The digitized material consisted of the Icelandic family Sagas and medieval literature preserved in manuscripts from NULI and the Árni Magnússon Institute and books about the Sagas and the manuscripts,

published before 1900, from the Fiske Collection of Cornell Library. The Fiske Collection holds the second largest collection of Icelandic literature outside Iceland, after the Royal Library in Copenhagen.

Implementation of the project had several major phases: design, programming, cataloging and digitization and all three institutions contributed, in varying degree, in each phase. Material from each institution had its own character, and therefore each institution faced different challenges, especially in digitization and cataloging. Programming involved challenges in developing software, which has now become outdated and the site is no longer in service. But fortunately it can be viewed in the Icelandic *Web Archive* (Vefsafn.is) using the URL Sagnanet.is as a search command.

Based on the work in *Sagnanet*, a new project has emerged, the *Manuscript Web* (Handrit.is) which is a union catalogue and a digital library of Old Icelandic and Norse manuscripts. This is also a cooperative project between NULI and the Árni Magnússon Institutes in Reykjavík and in Copenhagen and was sponsored by the eContentplus programme under the Enrich project (Driscoll and Haswell 2011, Hrafnkelsson 2011). These three institutions hold the major part of known manuscripts written in Iceland or Icelandic.

The *Manuscript Web* project serves two purposes. First, to convert old printed manuscript descriptions to detailed electronic format, enrich them with new information and make them easily available. Each institution had been experimenting with electronic cataloguing but using different formats. In NULI the manuscript descriptions for *Sagnanet* were in MARC, the other had used older version of TEI. All descriptions have now been converted to TEI P5 format.

The second purpose is to digitize all manuscripts preserved in those three institutions. All images from the older *Sagnanet* are reused and converted to a new format and made accessible on the site. New manuscripts descriptions and images are added on a regular basis.

The idea behind the *Manuscript Web* is to develop a research tool and it can only be achieved in close cooperation between the three partners. A project group meets on a regular basis to discuss cataloguing rules, methods for online access and searching capabilities. Now there are 7,200 descriptions of manuscripts accessible on the web and some are more detailed than others. Some 1,300 manuscripts have been digitized covering 336,000 pages. The project is ongoing.

NULI's biggest digitization effort so far is the *Newspaper and Periodical Digital Library* (Timarit.is). The project started in 1999 as collaboration between NULI, The National Library of the Faroe Islands and the National Library of Greenland. The project received a grant from Nordinfo and was sponsored by several parties in each of the participating countries, such as the Icelandic Research Council.

At the beginning, the aim was to digitize the collections of newspapers and periodicals from the three countries that were no longer in copyright (Hrafnkelsson 2001). The Icelandic material published before 1920 has already been digitized and made available and now material published 1921-1940 is being digitized. Many titles published in the 20th century have also been digitized and made available by contract with individual publishers. Major steppingstones were contracts with the biggest newspaper publishers, who sponsored the digitization of their own material.

Titles accessible on Timarit.is are now 742, covering 4.2 million pages. To help users access the material, Timarit.is has been linked to the national library system Gegnir.is and the national discovery service Leitir.is. The text of the scanned material was extracted using OCR technology and is searchable both on the site and in Google. Enabling searching via Google has substantially boosted the usage. The site is also frequently used for citations in Wikipedia. Timarit.is is by far the most popular collection of NULI (digital or otherwise) with over 15 thousand unique visitors each week. It is one of the thirty most used sites in Iceland (Modernus 2012).

A similar project, a *Digital Library for Books* (Bækur.is) opened for the public in December 2010. It is an ongoing project in several phases. In the first phase the oldest Icelandic print will be digitized, from the 16th century and ending in the year 1844. That year marks a moment in the printing history of Iceland, when the only operable printing press in the country was moved from the island Viðey to the town of Reykjavík and a new era of printing began.

The NULI collection of printed Icelandic books is not complete and cooperation with other libraries holding missing copies is therefore feasible in order to build a complete collection on the web. Total number of works to be digitized in the first phase is around 1,700. The next phase will be prints from 1845 to 1900. As the other projects, Bækur.is is connected to Gegnir.is and Leitir.is as well as being searchable via

Google. For the moment the book site is a work in progress. One can search for titles, authors and in the case of titles not printed in Gothic characters, make a free-text search.

Digitization of all the material mentioned above has been carried out at NULI. The digitization processing line, as it is commonly known in-house, is capable of scanning around 40 thousand pages per month. NULI operates digitization efforts in two locations, in Reykjavík and Akureyri, using Zeutschel scanners, with a total of 4.5 FTE's. The images are processed, including OCR extraction of the text before being reviewed again for accuracy. After quality check, they are pushed to the relevant websites as JPEGs and PDFs. The original TIFF scans are kept as well for preservation purposes.

Born Digital

Born digital material such as text, audio and video that has been produced and distributed as digital media has become an increasing concern at the NULI as in other national libraries. With the rise of the Internet a large part of formerly printed material is now only to be found online. Not only does this save money for the publisher, it also makes it much easier for the end user to find and consume. But different publishing methods and changing technology means change for national libraries, breaking the legal deposit setup that has ensured that published works find their way into their archives.

In 2002 the Icelandic Parliament, Althingi, extended the scope of the legal deposit legislation in order to include the new types of material. While it was possible to collect traditional media by imposing a deposit obligation on publishing and printing houses, a new approach was needed for the World Wide Web and electronic publishing. The law states that the only obligation placed on the publisher or maker of electronic works online, is that it should be made accessible to the depositor. NULI therefore had to reach out and collect the material (Hallgrímsson 2006b).

International cooperation turned out to be essential to enable NULI to acquire the skills and technology needed to meet this challenge. NULI had been working with the other Nordic national libraries on web archiving since 1999. Seven other national libraries joined the group and the International Internet Preservation Consortium (IIPC) was formed in 2003, to aid in the development of methods for collecting web based material. Through this cooperation a web harvester, named *Heritrix*, was developed and taken into use in Iceland (Sigurðsson 2005a).

The legal deposit legislation mandates NULI to collect all web based works that relate to Iceland or are made by Icelandic citizens. This means that the Library has to harvest immense amounts of data. Fortunately, most of it is contained within the national top level domain (ccTLD) for Iceland, *.is*. Using the newly built Heritrix web harvesting tool, the entire .is ccTLD was harvested for the first time in late 2004.

The first collection lasted for nearly two months and after considerable effort it generated around 710 gigabytes of compressed data (Sigurðsson 2006). At that time the .is ccTLD had about 15 thousand registered domains. Since then, these full domain harvests have been conducted three times a year. The size of the national domain has grown considerably and there are now nearly 39 thousand domains. A full scale crawl covers more than five thousand gigabytes before the data is de-duplicated and compressed.

While the full domain harvests form the foundation of the *Icelandic Web Archive* (Vefsafn.is) they are nevertheless incomplete. They do not capture data outside the national domain nor do they capture what happens between the harvests. To address the former concern, a curated list of relevant material hosted outside .is was made. This has proven to be extremely labor intensive and the list does only cover a small sample of material outside .is. One can foresee that this process can be automated to some extent in the future, but at present the appropriate tools are not available (Sigurðsson 2005b).

To capture data about events between crawls, an experiment of compiling a list of high value sites with content frequently updated was done in 2006. The list formed the basis for weekly harvests of a much more limited scope than the full domain harvests. The weekly harvests fill in some of the blanks between the large ones and ensure that more topical content is captured. There are expectations to extend this even further and collect on a daily basis or even more often. This type of harvesting, however, will always be highly targeted to few sites. It is impossible to collect the entire Icelandic web in week, let alone in a day. The list of high value sites has been updated regularly since 2006 and is maintained by the legal deposit unit of NULI.

To augment the weekly harvests, special harvests are made for special events (such as elections). They are organized like the weekly harvests, but use a list of sites tailored to the event and may run on a schedule other than weekly.

To date NULI has harvested over 34 thousand gigabytes of compressed data, visiting over 1.8 billion URLs. This covers all kinds of data, from simple text files, to PDF, audio and video files.

To access this vast amount of data the *Web Archive*, (Vefsafn.is) was opened in 2009. The site uses the Open Wayback Machine that was developed at The Internet Archive. Currently, it is only possible to look up data in the collection by its original URL. In the future full text searches may become possible (Hallgrímsson 2006c). All collected material is available to the public with some minor exceptions. Apart from the public, academics have also found ways to use the Web Archive. For example, texts from the archive have been used to build a frequency dictionary of the Icelandic language (Quasthoff, Fiedler and Hallsteinsdóttir 2012).

While the *Web Archive* is the most significant of the born digital collections at the NULI, it is not the only one. In recent years, as ever more content is published as digital-only, it has become convenient to establish a curated collection of material that is published as PDF or ePub, including e-books, reports, newsletters, web magazines and pamphlets of various characters. To handle this, *Digital Archive Rafhlaðan* (Rafhladan.is) was developed.

Unlike the web archive, *Rafhlaðan* will rely on publishers and individuals to voluntarily submit their data. Additional high value material is collected by staff for inclusion. This collection can be regarded as a response to the untamed and non-curated nature of the *Web Archive* where particular works can be difficult to locate. It is also a reflection of the ever growing trend towards digital only publications, notably e-books.

A part of the material in *Rafhlaðan* is subject to copyright and is not accessible to visitors of the site unless within the walls of NULI. Only the cataloging data or metadata is accessible in such instances. *Rafhlaðan* is a recent project and was opened in December 2011 and currently contains about 2,500 items.

Similar to *Rafhlaðan*, the Library also operates a digital repository for the universities in Iceland. The repository, named *Skemman* (skemman.is), was originally established at the University of Akureyri in 2007 and was moved to NULI when the University of Iceland became involved in the project in 2009. It contains academic and research works by students and staff of the universities.

Currently, there are over 10,000 works in the repository. The author of each work can determine the access policy. Most works are in open access, but access delays or embargos have been imposed on others. In some cases this can be as long as 120 years, although 3-5 years is more common.

Both *Rafhlaðan* and *Skemman* are based on *DSpace* repository software with some modifications.

Looking Forward

Digitization of Icelandic heritage preserved in NULI has been one of the main developing projects of the Library. It will continue for the next years but the progress depends on funding. It can be predicted that the main bulk of Icelandic published works in public domain will be in digital format and accessible on the web in 2018 when the Library will celebrate its 200 years anniversary.

Harvesting the .is domain will continue as well as collection of born digital material. The web archive will expand in the coming years and it will be interesting to monitor visits to the archive and discover how people will use it. As noted, academics have already discovered it as a potential research field.

Acquisition of born digital material will be the new developmental area for the next years. Definition of the works that will be collected is necessary and development of processes and ingestion services for the legal deposit partners will continue. Electronic legal deposit will grow considerably in the next years and NULI must be able to receive the material and store it for long time preservation. These tasks will move from development stage and become a part of everyday life in the Library. One can also foresee joint services on a national basis that are likely to emerge, such as a consortium of a national preservation repository. The goal would be to build a national trusted digital repository for long-term preservation of published digital material. Also to find new partners in preservation of digital material, i.e. radio, music business, archives, museums, films, photo archives etc. A national repository of scientific research and research data is another future vision that will probably be realized in the coming years (Sverrisdóttir 2009).

One of the most important aspects of this work is access and the use of all this material. How can it be opened and made available and usable for the general public? The easiest way is to expose the material to search engines and unified search portals. This requires the use of standardized metadata in order to be able to reuse, cooperate and link the data to other datasets, services and projects. The Library currently makes metadata from all its collections, except the *Web Archive*, available via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Material from NULI or connected to the Library is now searchable in Google and Google Scholar, Europeana, The European Library (TEL) and the Icelandic portal or discovery service Leitir.is. This has resulted in increased traffic on the websites and more use of the digital material.

Exposing the material in the right places is another important element regarding access. The social media, smartphones, e-readers and tablets is a new way to spread information. The first e-reader and smartphone edition of the web Bækur.is was released in May 2012, and other websites of NULI will follow. To appreciate the impact of social media it is worth noting that the second largest source of traffic for Timarit.is (after Google) is Facebook.

An important project that has made great effects within the library community and enhanced the services of libraries in Iceland is Gegnir, the national library system and online catalogue. It is operated by the Icelandic Library Consortium (Landskerfi bókasafna) which is a private company in public ownership, founded in 2001. About 300 libraries, information centers and other entities use the system. A large part of the cataloging in Gegnir is facilitated by NULI and the Library has responsibility for quality control of the data and the National Bibliography. In 2011 a new search machine, *Leitir.is*, was launched, building on Primo from ExLibris. Users can now search several databases from one platform and the idea is to build a national portal for digital cultural heritage, but also a search machine for licensed material. Available now is Gegnir (library catalogs), digitized books and serials from the NULI, collections from the Reykjavik Museum of Photography, two repositories with research and university related material, and hvar.is (licensed electronic journals and databases).

Many of the projects mentioned above required complex cooperation with many institutions and across national borders. Participation in international development teams like IIPC, TEL and Europeana has proven to be invaluable. Within NULI people have formed new kind of groups across boundaries. Programmers, librarians, catalogers, historians, archivists and web masters have worked together to realize a vision. This has resulted in extremely valuable services for the Icelandic community, services that enhance learning and culture and it is quite clear that a Digital National Library is emerging.

References

Driscoll, Matthew J. and Eric Haswell. 2011. "ENRICH: The way to seamless access on manuscripts and early printed books." *Care and conservation of manuscripts* 12: 1–9. Proceedings of the 12th international seminar on the care and conservation of manuscripts, University of Copenhagen, October 14–16, 2009.

Forn Íslandskort = Exploring old landscapes. 2000. Edited by Emilía Sigmarsdóttir, Jökull Sævarsson and Mark Cohagen. Reykjavík: Landsbókasafn Íslands – Háskólabókasafn. 19th International Symposium of the International Map Collector´s Society, Reykjavík, September 15–18, 2000.

Hallgrímsson, Þorsteinn. 2006a. "Sagnanet – SagaNet." *Nordisk tidskrift för bok- och bibliotekshistoria,* 6: 235–245.

Hallgrímsson, Þorsteinn. 2006b. "Web archiving, challenges and problems." In *KB, Kungliga biblioteket i Humlegården och i (cyber) världen*, redaktör Margareta Törngren, 381–403. Stockholm: Kungliga biblioteket.

Hallgrímsson, Þorsteinn. 2006c. "Access and Finding Aids."In *Web Archiving,* editor Julien Masanès, 131–151. Berlin: Springer.

Hannesdóttir, Sigrún Klara. 2005. "Library development in electronic environment: Iceland 2005." *IFLA Journal* 31 (2): 151–161.

Hrafnkelsson, Örn. 2001. "The VESTNORD project: digitizing newspapers and magazines from the 18th and 19th centuries." In *Nordiskt Forum for forskningsbibliotekschefer*, 131–138. Helsingfors: Nordinfo.

Hrafnkelsson, Örn. 2011. "The making of a manuscript digital library: The ÍBR collection past – present – future." *Care and conservation of manuscripts* 12: 11–19. Proceedings of the 12th international seminar on the care and conservation of manuscripts, University of Copenhagen, October 14–16, 2009.

Modernus. 2012. "Coordinated webmeasure." Accessed September 4. http://veflistinn.is/

Quasthoff, Uwe, Sabine Fiedler and Erla Hallsteinsdóttir. 2012. *Frequency Dictionary: Icelandic*. Leipzig, Leipziger Universitätsverlag. (Frequency Dictionaries; 3) http://wortschatz.uni-leipzig.de/ws_isl

Sigurðsson, Einar. 2000. "The next ten years in national libraries: the National and University Library of Iceland." *Alexandria*, 12 (2): 134–135.

Sigurðsson, Haraldur. 1971–1978. *Kortasaga Íslands*. Reykjavík, Menningarsjóður. 2 volumes.

Sigurðsson, Kristinn. 2005a. “Adaptive revisiting with Heritrix.” MS thesis, University of Iceland.

Sigurðsson, Kristinn. 2005b. “Incremental crawling with Heritrix.” Paper presented at 5th International Web Archiving Workshop (IWAW05), Vienna, September 22–23. http://www.iwaw.net/05/

Sigurðsson, Kristinn. 2006. “Managing duplicates across sequential crawls.” Paper presented at 6th International Web Archiving Workshop (IWAW06), Alicante, September 21–22. http://www.iwaw.net/06/

Sverrisdóttir, Ingibjörg Steinunn. 2009. “The next ten years in national libraries: The National and University Library of Iceland.” *Alexandria,* 21 (2): 57–58.

The NULI websites referred to in the text usually have an English version or information in English

Islandskort.is	Antique maps
Handrit.is	Manuscripts
Timarit.is	Newspapers and Periodicals
Baekur.is	Books
Vefsafn.is	Web Archive
Rafhladan.is	Legal Deposit - born digital material
Skemman.is	Digital repository for universities
Gegnir.is	National library system and a national bibliography
Leitir.is	National digital portal or discovery service
hvar.is	Iceland Consortium for Electronic Subscriptions (ICES)