CORPORA, ELIBRARIES AND DATABASES: LOCATING SYRIAC STUDIES IN THE 21ST CENTURY*

KRISTIAN S. HEAL BRIGHAM YOUNG UNIVERSITY

ABSTRACT

This paper examines the implications of current library, publishing and research trends for Syriac studies. A good research library is essential to good scholarship, but the medium for delivering that library is changing, and Syriac scholars are increasingly working with both print and digital resources. Specialist on-line collections are supplementing the holdings of massive on-line repositories, providing open access to the majority of out-of-copyright works in the field. Moreover, it is becoming increasingly clear that the printed book is no longer the ideal container for certain types of information, and digital corpora and databases are being created and conceived to replace their print equivalents. Thus, databases are emerging as dynamic replacements for traditional Syriac reference works, and digital corpora are in preparation that supplement and facilitate access to the printed corpus of Syriac literature. In addition to examining the advantages promised to Syriac scholars by the digital research environment, this paper also considers some of the obstacles and disadvantages posed by this transition.

^{*} I gratefully acknowledge valuable conversations with and suggestions from Chris Erickson and Ryan Combs at BYU's Harold B. Lee Library, David L. Armond at the J. Reuben Clark Law Library, and Carl Griffin, my colleague at the Neal A. Maxwell Institute. I also want to pay tribute to Dr. George Kiraz, who has executed pioneering projects in all of the areas discussed in this paper.

To locate Syriac studies in the 21st century is to describe a research environment in which scholars work entirely without the printed page. This is not intended as a provocative or dystopian notion—rather, it reflects the real and important shifts already taking place in the thinking and practice of library and publishing professionals, as well as in the way students and scholars are accessing texts and information. Since it is futile to resist the forces that are driving the digital future, the Syriac scholar would do well to be informed, to make use of the emerging resources, and be involved in shaping the tools that will best enhance the way we work and open the field to a wider variety of researchers. It is also important to recognize the limitations as well as the benefits of this new research environment.

LIBRARIES AND ELIBRARIES

The history of Syriac studies could well be told in terms of the scholars' search for, or frustrated separation from, the books they wish to study. Hunayn ibn Ishaq, for example, sounds surprisingly contemporary when he writes in his letter on his Galen translations about the "books [he] had gathered one by one, during the whole of [his] life, since the time [he] had come of age, from the various countries in which [he] had traveled." Likewise in the numerous letters written by Timothy I to locate particular Syriac works, we recognize the kind of requests that today are daily directed to the *Hugoye* email list. The problem suggested by these letters from the past and their equivalents in the present is that only in very rare circumstances are the scholars and the books in the same place. There are Syriac libraries without scholars and well-trained Syriac scholars with no library. This problem is exacerbated by the growth

¹ A useful discussion of the feasibility (as of 2010) of an all-digital research library can be found in Lisa Spiro and Geneva Henry, "Can a New Research Library Be All-Digital," in *The Idea of Order: Transforming Research Collections for 21st Century Scholarship* (Washington D.C.: Council on Library and Information Systems, 2010), 5-80.

² John Lamoreaux, *Hunayn b. Ishaq: On his Galen Translations* (Provo: Brigham Young University Press, forthcoming) §2.4. Books which he then "lost in one fell swoop"!

³ They are letters 3, 16-20, 22, 24, 33, 37-39, 43, 47, 49, as mentioned in Sebastian P. Brock, *A Brief Outline of Syriac Literature* (Kottayam, India: SEERI, 1997), 64.

of the field in the Middle East, India, and North America, where students trained in the great universities are getting jobs and making careers in far less well-provisioned circumstances.

Building a Syriac studies research collection has until now been the answer for an institution that wants to support a scholar working in this field. However, since collection development is not simply a matter of a library buying books, but also of cataloguing, accessioning, maintaining and curating a collection, building such a collection represents a substantial long-term commitment.⁴ Many institutions will simply want to rely on inter-library loans to deal with the more arcane requests, though this is not inexpensive, and such institutions eventually lose credibility as net-borrowers in the system. Most importantly, however, is the fact that even in the best of circumstances where a Syriac scholar is fully supported in their research by their library, such print acquisitions support only a single scholar. When that scholar moves or retires, a Syriac specialist is not likely to take their place (because their college generally wasn't looking for a Syriac specialist in the first place), leaving the college with a collection that it still has to pay to maintain. Libraries are conflicted in situations like this. They want to serve patrons, while at the same time responding to the exigencies of space, costs, efficiency, competition, new technologies and changing metrics of effectiveness.

Of course the greatest change in the world of library science is the relatively recent rise of massive online book and document collections. Many online collections contain aggregated content

⁴ See Paul N. Courant and Matthew "Buzzy" Nielsen, "On the Cost of Keeping a Book," in *The Idea of Order: Transforming Research Collections for 21st Century Scholarship* (Washington, D.C.: Council on Library and Information Resources, 2010), 81-105. Even if Courant and Nielsen's estimates for the cost of maintaining printed books are inflated and the cost of maintaining digital resources naively low, as the question becomes one of either print or digital, print will lose out not least because of changing library usage practices. Note, for example, how the high estimated costs by Courant and Nielsen prompted Steve Kolowich to observe that, "The administrators who provide library budgets may be reluctant to fund new facilities to house print collections and may question large expenditures to support both print and electronic formats. Library directors must consider not only the immediate expectations of faculty, but also the long-term goals for the library" ("E-Library Economics." *Inside Higher Ed.* February 10, 2010).

produced by library services companies, accessible to libraries by subscription.⁵ Individual publishers, such as Cambridge, Oxford and Brill also have their own online subscription services. Through these services, libraries and their patrons have access to hundreds of thousands of electronic books, journal issues and dissertations from a variety of sources.

Other projects are generating or providing freely available digital content, mostly out of copyright material. The *Google Books* project has scanned over ten million volumes in only five years. *Europeana* aggregates content from all over Europe, making over six million digital items available. *Internet Archive* offers over a million volumes, in addition to millions of other digital items. What's more, *Internet Archive* have now relaunched their Open Library project in an effort to use crowd-sourcing to create an online annotated catalogue of every book ever published, linking out, wherever possible, to electronic copies in its own holdings. This is to say nothing of the thousands of smaller online collections and websites that support the work of hundreds of academic disciplines, including Syriac Studies.

Several institutions have created online collections for Syriac studies. I mention here only the four collections with which I am best familiar, namely the joint BYU-CUA Syriac Studies Reference Library (online since 2007); the joint Beth Mardutho-CUA eBeth Arké, now hosted by HMML; the valuable collection recently put online by the Goussen Library in Bonn; and the collection of books, articles and manuscripts available on the website of the Center for the Preservation of Ancient Religious Texts at BYU (see the Appendix for details). These institutional efforts are

⁵ Such as *Ebrary* (more than 435,000 titles), *JSTOR* (more than 260,000 journal issues), *Project Muse* (more than 144,000 journal articles), *Questia* (more than 76,000 books and over 2 million journal articles) and *Proquest* (more than 2 million dissertations).

⁶ http://openlibrary.org/ (last accessed 5/26/2011).

⁷ http://www.intute.ac.uk/ (last accessed 5/26/11) provides an annotated catalogue of 120,000 websites useful to students and scholars, including over 1,400 for religion and theology. Note, however that funding for the maintenance and development of this site ended in July 2011. Also, see Dr. Meriel Patrick's useful article, "Disentangling the Web: A Guide to Online Resources for Theology," *The Expository Times* 121.5 (2010): 213-217.

supplemented by an informal network of colleagues exchanging copies of electronic books they have found online or have produced themselves.

Most of these projects started before the Google Books initiative, and despite the massive amount of work done by Google, these specialist collections still contain many unique items. This is not just because Google has digitized less than 10% of the 125 million volumes recorded in WorldCat. Rather, it is because many books in our field are only found by those who, like Hunayn ibn Ishaq, are willing to spend a lifetime searching for them in the dark corners of the world's libraries, bookshops and personal collections.

Of course, Hunayn's frustrations are read with particular empathy by any scholar who has tried to find, and then see a manuscript held in a Middle Eastern collection. This is why, without doubt, the most important digital initiative in our field in the past few years is that of the Hill Museum and Manuscript Library, under the guidance of Fr. Columba Stewart.8 However, although it is hoped that this project will ultimately result in an online collection, patrons not working in Collegeville will still receive their manuscripts through the post, though on a disc instead of on microfilm. The prize for the largest online collection of manuscripts may yet go to the Vatican Library. However, though they have announced plans to undertake a ten year project to digitize their entire collection of 80,000 manuscripts, it is not yet clear how and under what terms it will be made available. We can only hope that the arrangement will be a success, and encourage other major research libraries to do the same with their manuscripts.

The point of this recitation is twofold. Firstly, to show how quickly we can transition into an era where scholars in Syria or Saskatchewan have online access to a greater collection of books and manuscripts in Syriac studies than they could ever hope to find in any single library. Secondly, these various projects and services indicate that libraries are quickly entering this new world of digital content. An inherent conservatism may have led to a substantial overlap between the print and digital holdings in most libraries, but

⁸ Two other projects, which we all hope will one day end up as online collections, make up in quality what they may lack in quantity, namely the exciting new finds of Deir es Suryan, and St. Catherine's Monastery.

the difficult decisions forced upon libraries by the economics of the digital revolution are no longer being deferred. There are also still concerns about the possible incompatibility of the goals and objectives of libraries and the various digital content providers, and certain copyright issues may also be unresolved (so none are willing to adopt a scan and burn, or digitize and discard policy). Yet, while these issues are being resolved libraries are ineluctably transforming into subscribers, creators and curators of digital collections. Moreover, libraries will increasingly be assessed not simply by the number of books on their shelves and the number of people who walk through the door, but also by the number of bytes they have produced in online collections and the number of hits on their website.9 Thus libraries are turning to their own unique content and making that material the focus of new online collections. This trend has and hopefully will continue to benefit Syriac studies directly.

DATABASES AND THEIR USES

Many scholars exhibit the propensity to collect and organize materials, to produce finding aids and other such data and metadata. This quality is not always fully appreciated. The more dynamic C.C. Torrey, for example, is said to have thought Carl Brockelmann dull because he compiled "vast files of notes on cards." Of course, the value of Brockelmann's early databases is demonstrated by the daily recourse made by scholars to the works derived from them (Torrey's work, on the other hand, has not aged quite so well). Anton Baumstark's invaluable *Geschichte der syrischen Literatur* is likewise the product of an inveterate collector of data. 11

⁹ Since 2006, the Association of Research Libraries (ARL) has surveyed libraries about the collections they have digitized, the number of items in that collection, and the total size of digitized collections. See http://www.arl.org/stats/annualsurveys/sup/index.shtml (last accessed 5/26/2011). I'm grateful to David Armond for this note.

¹⁰ Cyrus Gordon, "Philology of the Ancient Near East: My Seventy Years in Semitic Linguistics," *Built on Solid Rock. Studies in Honour of Professor Ebbe Egede Knudsen on the Occasion of his 65th Birthday*, ed. E. Wardini (Oslo: Novus Verlag, 1997), 91-101, citing from page 97.

¹¹ In a current project to prepare an index to his manuscript citations, for example, we have discovered that Baumstark makes no less than nine and half thousand individual references to Syriac and other manuscripts.

It may seem strange to talk about Baumstark and Brockelmann creating databases, until we recognize the fact that the handbook, the history of literature, the bibliography, the clavis, the manuscript catalogue, the concordance and the lexicon are all, in essence, databases. The problem is they are databases that have become calcified in print, leaving us with only a limited number of queries that we can make of the data, queries that had to be anticipated by the author and added to a print volume in the form of indices and tables or other structural features. Though such indices and structural features are a form of expert filtering, as a digital resource such filtering is not as static as it is in print publications. You can have a copy of Jessie Payne Smith's dictionary that can be displayed in both root and alphabetical order, in which the usage examples can be searched as well as the headwords—a dictionary that works just as well from English into Syriac as it does from Syriac into English. When converted into electronic databases, you can easily look up the manuscript citations in Baumstark; view Moss by ancient or modern author, or by subject, or year of publication, or publisher or journal; you can examine manuscript descriptions in chronological order, or by scribe, or monastery, or donor, or original collection, or by content, or any combination of these—it is all a matter of how the data is tagged. Put in these terms, hopefully it is clear not only that we are all already using databases, but also that our existing and future tools should be accessible as electronic databases even if they also appear and are used as print publications.12

In preparing a database, careful consideration needs to be given to encoding or tagging the data, or setting up the right fields in a database structure. In the end, a database is only as useful as its structure permits. Usability is a function of filtering the data and the better the data is encoded the more effectively it can be filtered. The TEI standards have been well thought out in this regard, and worth adopting when taking a standards-based approach to preparing and annotated a database. However, as many standards have been created for projects much larger and more complex, it is

¹² It is vital, however, not to discard the achievements of such print resources when converting them to a digital format. The presentation layer or user interface of a database should be every bit as elegant, well-designed, typographically pleasing and easy to use as the best print publications.

important not to let the over engineering impulse drown a project in metadata.

Perhaps the single most pressing concern regarding database projects is the risk of obsolescence. A database can be lost because of its dependence on obsolete software or hardware. For data to survive it needs to be standards based, platform independent and regularly migrated and updated—in other words, it requires a lot more care and feeding than a book! The most effective method of ensuring the longevity and use of data is to share it. In the library world they use the acronym LOCKSS—Lots of Copies Keeps Stuff Safe. This applies to all digital materials, not only databases. Unless a database is maintained by a business or an endowed Institution it will almost certainly become obsolete in a generation, if not earlier. However, if it is constructed using standardized encoding, such as TEI, and set free to be repurposed and reused in a variety of different environments, it may long outlast its original creator.¹³

All of these considerations are evident in the extremely well executed and promising *Syriac Reference Portal* that is being built by the Syriac Research Group at the University of Alabama under the direction of David Michelson. The aspirations of this project are the natural extension of a full appreciation of the power of well-structured data. In this case, the combination of structured manuscript catalogue data, classified bibliographies, geographic data, and a muliti-lingual authority file for referencing Syriac authors, works and place names (supplanting a traditional Clavis and gazateer) will result not only in a powerful reference tool but in a new ontology of Syriac Studies.¹⁴

The Comprehensive Bibliography on Syriac Christianity is another database that will doubtless become foundational to the digital infrastructure of Syriac Studies, not only because it comprises over 14,000 bibliographical entries, but because these entries have been enriched with more than 2,500 different keywords to facilitate an

¹³ "Resources that are encoded using open standards have a greater chance of remaining accessible after a long time, rather than those resources that are not." Herna L. Victor et. al, "Preserving Object-relational Databases for the Next Generations" (Fifth International Conference on Digital Information Management, 2010), 12. I thank David Armond for this reference.

¹⁴ http://www.syriac.ua.edu/ (last accessed 5/16/2011)

array of specific and general searches. Like all such tools, ¹⁵ this *Comprehensive Bibliography* builds upon decades, if not centuries, of patient bibliographical research. However, the new medium of delivery has allowed the editor, Sergey Minov, to think beyond the temporal and disciplinary boundaries that define all existing print bibliographies and create something that is deliberately comprehensive, not only in terms of coverage but also access.

CORPORA

To call an annotated corpus a database, though accurate, is perhaps misleading. It may be better to think of a corpus as structured and annotated text. The purpose of adding structure and annotation is not to bypass engagement with the text but to facilitate it—to expedite the space between ideas and engagement with the textual evidence related to those ideas. Certainly this is the goal of the BYU-Oxford Corpus of Syriac Literature.

When this corpus project began five years ago, we felt that it was best to begin with a small feasibility study. This was to involve preparing a modest sample corpus, and working with computer scientists to demonstrate that with the right computational framework we could indeed greatly reduce the time it traditionally took to annotate a corpus. The problem we wanted to solve was articulated by George Kiraz nearly twenty years ago when he observed, "The major task in concordance generation is the tagging of each word in the text... This [...] is a long and tedious process. If one considers generating a concordance to the OT or the works of Ephrem, this process can take years, if not decades (it took the Way International 15 years to produce their database of the NT!)."¹⁶ It was simply necessary to find a better way, hence the crucial collaboration with computer scientists.

Thus far, over six million words of Syriac have been transcribed from published editions and manuscripts, representing over 500 individual works in a variety of genres, written by nearly 100 different authors spanning the second to the eighteenth

¹⁵ For example, the Hebrew University's RAMBI database (http://www.jnul.huji.ac.il/rambi/).

¹⁶ Kiraz, George A., "Automatic Concordance Generation of Syriac Texts," in *VI Symposium Syriacum*, ed. René Lavenant (OCA 247. Rome: Pontifical Oriental Institute, 1994), 461-75; citing from page 475.

century.¹⁷ An electronic lexicon is in preparation, which will be integrated into the corpus and also serve as a stand-alone resource. Five years of research has been invested in developing efficient computational tools for segmenting and tagging the corpus. Major computational issues have been explored and resolved in published research papers, as well as dissertations and theses.¹⁸ The general progress of the project has also been presented at conferences, and the corpus is already informing the research of Syriac scholars.

The next step in the project is to begin to encode the texts according to TEI standards, and then proof, tag, and publish the corpus. The project is currently focused on texts that are part of specific research projects, including a projected concordance to the poetic works of Ephrem the Syrian. However, even in its raw state the corpus has some value to scholars. Hence the plan to begin to publish the entire corpus immediately text by text as the TEI encoding is added (these texts will be labeled as provisional). Then, as each text is proofed and corrected, and each word is tagged and linked to the dictionary, it will be labeled as final data.

As the corpus grows the lexical database will expand. We anticipate adding numerous new words to the lexicon, as well as providing citations for rare words. However, these developments only hint at the scope and potential of a dictionary linked to an expanding corpus. Usage examples and frequency data will also be added to the entries, as well as details of the first occurrence in the corpus of a given word and its usage over time. We are thus creating the infrastructure in which a new and complete *Thesaurus Syriacus* is finally conceivable.¹⁹

¹⁷ Equivalent to about 250 CSCO volumes (only 120 volumes of Syriac text have appeared in that series so far). Sebastian Brock once observed in casual conversation that, "The corpus is our life," referring to the CSCO. Certainly, CSCO, PO and other great Syriac editing projects stand at the heart of our corpus project. Like Jacques Paul Migne two centuries ago, we can only hope to do great things by building on the work of others—though unlike Migne we will scrupulously acknowledge our sources!

¹⁸ The research team has included three BYU faculty members from Computer Science and Computational Linguistics, three BYU PhD students, and two BYU MSc students, as well as several undergraduates.

¹⁹ It is our hope that the corpus project will be seen as a useful and integral part of many research projects, and that colleagues working on particular texts, genres, or authors will want their material included in the

While the Oxford-BYU Corpus struggles to become a viable research tool, scholars have ready access to the wealth of material available at the Comprehensive Aramaic Lexicon.²⁰ This integrated lexicon and corpus offers full keyword searches of a substantial collection of fully annotated Syriac texts, together with comparative lexical analysis.

CONCLUSION

Well-conceived digital resources clearly add value to the discipline, but they do not always have academic value according to the metrics currently used for rank advancement or assessment exercises. For the time being this will inhibit their proliferation. This is a transitional period, one in which scholars working on corpus or database-like projects will need to take the initiative, and the risk, to think beyond their print publications, and bear in mind that the material that they have worked hard to gather and produce could be considerably more useful to a much greater constituency in the form of an electronic database. Thus, suitable projects should be undertaken with one eye on the present and one on the future.

However, the impermanence of digital projects is disconcerting. All web-based projects are currently hot-house flowers, requiring constant tending to survive. Close adherence to broadly accepted standards can help to mitigate such a concern, but what is needed is the digital equivalent of printing and binding—some more permanent receptacle for the distribution and preservation of data—before non-institutional databases can survive and flourish.

Finally, despite the triumphal tone of much of this paper, it is not yet clear how and to what extent the digital revolution will change Syriac studies. Certainly it will shorten the path from posing research questions to gathering the data to answer them. It could also help decrease the time it takes to obtain mastery in a research area. As R. Keith Sawyer observed, "sociocultural theory argues that before becoming creative, individuals must become socialized into the field and internalize the domain. In this approach,

corpus. However, we are also committed to growing and refining the corpus through our own efforts.

²⁰ http://cal1.cn.huc.edu/ (accessed on 11/15/2011).

individuals would not be expected to be productive or to generate important works until they'd fully internalized the domain."²¹ Digital resources could certainly help shorten the time it takes to internalize the field of Syriac studies and become a productive scholar within the field.

Will such benefits necessarily result in a steady democratization of the discipline? Certainly the ready availability of digital resources will enable Syriac scholars to flourish in any location, thus challenging the assumption that there is a necessary connection between our field and its historical centers of activity with their well-stocked libraries. Nevertheless, the well-stocked library of a great university is only one part of what generates a synergistic academic environment. Scholarship flourishes when one is immersed in an academic culture, and despite the rise of academic social media sites, on-line bibliographical guides and specialist mailing lists, there is still no digital substitute for the noetic connections made in the common room over a hot drink and a biscuit.

APPENDIX: SELECT DIGITAL RESOURCES FOR SYRIAC STUDIES²²

Books

CPART, BYU: http://cpart.byu.edu/?page=99&sidebar eBeth Arké, Beth Mardutho/CUA/HMML:

http://www.hmml.org/vivarium/BethArke.htm

Europeana: www.europeana.eu

Gallica: http://gallica.bnf.fr/?lang=EN Google Books: http://books.google.com/

Goussen library collection:

http://s2w.hbz-nrw.de/ulbbn/nav/classification/16431

²¹ Sawyer, R. Keith, *Explaining Creativity: The Science of Human Innovation* (Oxford: Oxford University Press, 2006), 162. I am grateful to David Armond for this insightful reference.

²² I'm grateful to Gabriel Rabo, Nikolai Seleznyov and Hidemi Takahashi for contributing to this list. Gareth Hughes (University of Oxford) has created a useful tool to simultaneously search all of these resources, which is available here:

http://www.garzo.co.uk/documents/Select_Syriac_eResources.html

Internet Archive: www.archive.org Ishtar Broadcasting Corporation:

http://www.ishtartv.com/books.html

Leiden University Repository: https://openaccess.leidenuniv.nl/ Suryoyo Online, Library:

http://www.suryoyo.uni-goettingen.de/library/

Syriac Studies Reference Library, BYU/CUA:

http://www.lib.byu.edu/dlib/cua/

Corpora

Comprehensive Aramaic Lexicon: http://cal1.cn.huc.edu/ Oxford-BYU Corpus:

http://cpart.byu.edu/?page=112&sidebar

Syriac Library (maintained by Roger Pearse):

http://www.tertullian.org/rpearse/thesyriaclibrary/

Databases

Comprehensive Bibliography of Syriac Studies:

http://www.csc.huji.ac.il/db/browse.aspx?db=SB

Dukhrana Multiple Lexicon Search:

http://dukhrana.com/lexicon/search.php

Syriac Reference Portal: http://www.syriac.ua.edu/

Email Lists

Hugoye List: http://groups.yahoo.com/group/hugoye-list/

North American Christian Arabic Society List:

http://groups.google.com/group/nascas

Suryoyo-Online:

https://listserv.gwdg.de/mailman/listinfo/suryoyo-online

Journals

Collectanea Christiana Orientalia:

http://www.uco.es/investiga/grupos/hum380/collectanea/node/4

Hugoye: Journal of Syriac Studies:

http://www.bethmardutho.org/index.php/hugoye/volume-index.html

Parole de l'Orient:

http://documents.irevues.inist.fr/handle/2042/34760

Syria. Archéologie, Art et histoire:

http://www.persee.fr/web/revues/home/prescript/revue/syria

Zeitschrift der Deutschen Morgenländischen Gesellschaft:

http://menadoc.bibliothek.unihalle.de/dmg/periodical/titleinf o/2327

Zeitschrift für Semitistik und verwandte Gebiete:

http://menadoc.bibliothek.unihalle.de/dmg/periodical/titleinf o/2341

Manuscripts

Beinecke Library, Yale:

http://beinecke.library.yale.edu/digitallibrary/

Biblioteca Medicea Laurenziana:

http://teca.bmlonline.it/TecaRicerca/index.jsp

CPART, BYU (St. Mark's Convent):

http://cpart.byu.edu/?page=126&sidebar

e-Corpus: http://www.e-corpus.org

Hill Museum and Manuscript Library:

http://www.hmml.org/vivarium/

Manuscripts of the Mediterranean (Manumed):

http://data.manumed.org/index.php

Mingana Manuscripts:

http://vmr.bham.ac.uk/Collections/Mingana/part/Syriac/

Various Syriac Manuscripts online (maintained by Steven Ring): http://www.syriac.talktalk.net/On-line-Syriac-MSS.html