Accessible Archives Responds to Our Customers’ Needs!

Carolina Consortium
Accessible Archives is pleased to join the participating academic and public libraries in the Carolina Consortium! We recently attended the Carolina Consortium Conference and Iris Hanney conducted a successful Premier on Accessible Archives!

Michigan & Pennsylvania County Histories Now Available
Accessible Archives announces the completion of Pennsylvania & Michigan in our landmark American County Histories Series. Accessible Archives is the only publisher that has collected and digitized all of the county histories of the U.S. – all 50 states and the District of Columbia in one database! We offer free MARC records, images and full text of all the books – over a million pages of content!

Expanded Direct Product Links
Accessible Archives has responded to requests from our customers for expanded direct browsing and search links for two of our most popular digital collections – African American Newspapers and American County Histories! Accessible Archives recognizes the value of these expanded links for use in a library’s research guides and libguides.

Achieving Higher Customer Satisfaction Is Our Goal at Accessible Archives

Katherine Brown, Collections Analyst, Aurania Library -- “Thank you so much for your help with figuring this out! I really appreciate your prompt responses and dedication to figuring out the problem.”

Elizabeth J. Cronin, Coordinator Information Services, Ocean County Library -- “The Military Newspapers of the WWI archive has been great to promote since it includes the Camp Dix paper. The picture of the Camp library is a treasure.”

Barbara Kelly, Director of Libraries, Faulkner University -- “Thank you so much! Thank you for working with us in the way that you have. I have to say, I have never had a vendor work with us so well. We look forward to continuing our patronage with you and marketing the product a bit to our students.” Barbara Kelly, Director of Libraries, Faulkner University

Angie Thompson, Cataloging Assistant, Liberty University -- “I really appreciate your quick response and timely resolution. I deal with a lot of our electronic content vendors when problems arise, and your team’s support is head and shoulders above the rest!”

Upcoming Conference Events
Will you be at the ALA Annual Conference, June 20-25, 2019? We’d love to visit with you at Booth 3041!
Contact us for an appointment; we have lots to talk about!

Walter E. Washington Convention Center, Washington, D.C.
Accessible Archives, Booth #3041

© 2019 Accessible Archives, Inc.
Accessible Archives is the exclusive sales and marketing agent for Accessible Archives:

Unlimited Priorities LLC is the exclusive sales and marketing agent for Accessible Archives:
Iris L. Hanney, President
Unlimited Priorities LLC
239-549-2384
iris.hanney@unlimitedpriorities.com
www.unlimitedpriorities.com

Robert Lester, Product Development
Unlimited Priorities LLC
203-527-3739
robert.lester@unlimitedpriorities.com
www.unlimitedpriorities.com

Publisher and editor of Inside the Archives

SUMMER 2019
VOLUME VIII, ISSUE 2

Text Mining in the Humanities: What’s a Library to Do?

So, are your humanists asking you for text mining support yet?
Some Humanities scholars have been engaged with these activities for many, many years now, but more and more seem to be getting involved as interest grows and tools become more user-friendly. Libraries increasingly face questions of how to appropriately support text mining activities. Here is a primer and pragmatic advice for those who are considering just how to do this.

Darby Orcutt is Assistant Head, Collections & Research Strategy, NC State University Libraries, Faculty, University Honors Program, Affiliated Faculty, Center for Innovative Management Studies, Affiliated Faculty, Genetic Engineering & Society Center, and Affiliated Faculty, Leadership in Public Science Cluster, as well as recently served as the Associate Chair of the Faculty of NC State. A national leader in developing models for access to proprietary and use-limited data for content mining and computational research, his current work revolves primarily around research support and engagement for interdisciplinary teams.

First off, what is text mining? Generally, the term is used to refer to a host of computational research practices where a computer “reads” texts at scale and uses algorithms (or artificial intelligence) to quantify or classify texts or their elements. Text mining results can expose linguistic or semantic features, recurrent correlations, and complex patterns across a large corpus. Patterns that a highly observant human scholar might recognize only after a lifetime of reading within their field - or never even notice at all - might be discovered by means of text mining within mere seconds.

Most librarians associate text mining primarily with the digital humanities, but text mining has been an established part of science and social science fields for a very long time. It has perhaps come to the awareness of most academic librarians because humanist scholars have started asking their libraries for help in ways that their more technologically- and quantitatively versed users may not have. In addition, as perhaps disproportionately heavy users of library resources, humanists naturally look to their library as a primary means of discovering information content, accessing it, and for help in using it.
If you see these activities - discovering, accessing, and supporting the use of information - as mission-critical for academic libraries (and I hope you do!), then you’re probably already involved in or at least planning for some effort to provide these to humanists around text mining. As with all areas, an institution needs to figure out what its user base is for the service at hand. Text mining currently represents a new interest for many scholars, who may need very basic support, but ranges all the way up to the highly knowledgeable and technically skilled (or humanists who partner with technical experts). When assessing your campus’s needs around text mining, understand who are approaching the libraries for help may represent only the more novice portions of your users who are engaged in text mining, and that you may need to seek out those who could benefit most from much of what you could offer.

Access

Access to content for text mining is perhaps the most fundamental - and most difficult - aspect of library support for text mining. As with all areas, an institution needs to figure out what its user base is for the service. If you see these activities - discovering, accessing, and supporting the use of information - as mission-critical for academic libraries (and I hope you do!), then you’re probably already involved in or at least planning for some effort to provide these to humanists around text mining. While we provide all of this wonderful content (databases, ebooks, electronic archives, and more), all of our collections have been built historically for human readers. For computers to “read” our electronic resources means that they have to be able to download (or at least somehow obtain) very large amounts of content, something that might not only overload the servers of the content provider, but also certainly violates traditional electronic resource contracts, which of course were designed with human readers in mind. Fortunately, excellent historical, literary, and cultural content for many eras and geographies can be found, particularly covering works that are no longer in copyright. Quite a few Open Access (OA) resources offer content through APIs (special online interfaces by which data can be easily downloaded en masse) or can be readily “scrapped” by researchers (“web scraping” is the technical process of automatically pulling data off of a web site and creating a structured data set from it). One of the more popular OA resources that provides ready access to digital humanities content is the Making of America (MOA) project, which was a collaboration between the University of Michigan and Cornell University, originally funded by the Andrew W. Mellon Foundation in 1995 (https://quod.lib.umich.edu/m/moagrp/index.html).

In terms of licensed and proprietary content, it has become much more common in recent years for vendors to offer data access, although this is still not a standard contract term for most academic libraries (libraries who are licensing new resources, as well as by seeking addenda on existing resource contracts). The time to ensure access to your collections for text mining purposes is prior to the need, as these licensing processes can take some time, even when the parties readily agree to business terms. For digital collections that are available in library, then some reasonable means of data access for purposes of computational research should be included; libraries should avoid investing in content whose use is contractually limited to non-computational readership. For leased collections, the marketplace has not yet fully settled on one model, for the very pragmatic reason that it’s not necessarily reasonable for a text mining project (which may last for many years) to continue to use robust data sets beyond the period of leased access.

Accessible Archives was one of the first commercial vendors of historical archives to offer equal access for text mining as for human readers for perpetual access content. A data mining addendum is included in their standard license agreement. In addition, as Accessible Archives prides itself on the high quality of its TEI Lite XML and rekeyed content at 98% quality or better, their collections more easily suit the technical needs of many especially mid-range humanities researchers. High-end data science researchers can handle content in virtually any way that they can get it, although well-structured metadata is usually preferred. Technically less sophisticated researchers (which are the vast majority of humanist text miners) generally require fairly well-structured metadata to accomplish their work.

Discovery

Researchers who engage in text mining often have great difficulty in finding accessible data sets. Even libraries that have worked hard to make their collections more user-friendly do not yet represent them well - in part, because as a research library community, the frequently thorny issues of library catalogs and usually only via lists on library web sites of resources that can be mined by authorized users. This effectively means that researchers who look for resources first within the silo of the type of research they wish to do (text mining) and only then based on the nature (subject, period, genre) of the content itself. Clearly representing the means of accessing data sets also proves challenging, as these are quite diverse, ranging from APIs to mediated vendor requests via librarians to even local library storage on hard drives. In an ideal environment, libraries would readily offer samples of the data set as well so that researchers can make sure that they have the capacity to deal with any particularities of its formatting. Lastly, the provenance and history of the data set (if even known by its user) may be vital to a given researcher, especially as metadata practices may have changed during the course of its digital production, particularly for resources that were created or revised over a longer period of time. Believe it or not, there are only some of the aspects of a data set that may be crucial to a researcher in finding an appropriate corpus to mine.

Services

Providing appropriate services in support of text mining activities is similarly challenging, and should be very context-driven, reflecting the user needs, mission, priorities, and capacity of the individual library. For high-end researchers, simple access and discovery support may be adequate, as they already have the tools, expertise, and support structures to conduct their research. Yet, the majority of our text mining users (and the fastest growing demographic) fall somewhere in the range between novice and knowledgeable non-expert. Scaling support services to this community on your campus requires knowing that community and recognizing that its needs may change rapidly.

Will your library support text mining tools? Many strong tools and platforms for text mining (which may be available on your computer) may refer users; one of the more venerable sets of textual analysis tools is MALLET (Machine Learning for Language Toolkit), produced by the University of Massachusetts at Amherst and freely available online at http://mallet.cs.umass.edu/. Many more or less “out-of-the-box” applications satisfy the needs of novice or occasional textual analysts, and easy web searches reveal the most popular tools that other librarians recommend and promote to their users, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization tools, and a host of other text mining related subjects to users hungry for this content. If you are wondering whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information services provider, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library support text mining tools? Many strong tools and platforms for text mining (which may be available on your computer) may refer users; one of the more venerable sets of textual analysis tools is MALLET (Machine Learning for Language Toolkit), produced by the University of Massachusetts at Amherst and freely available online at http://mallet.cs.umass.edu/. Many more or less “out-of-the-box” applications satisfy the needs of novice or occasional textual analysts, and easy web searches reveal the most popular tools that other librarians recommend and promote to their users, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization tools, and a host of other text mining related subjects to users hungry for this content. If you are wondering whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information services provider, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization tools, and a host of other text mining related subjects to users hungry for this content. If you are wondering whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information services provider, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization tools, and a host of other text mining related subjects to users hungry for this content. If you are wondering whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information services provider, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization tools, and a host of other text mining related subjects to users hungry for this content. If you are wondering whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information services provider, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on library computers will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.
Access to content for text mining is perhaps the most fundamental - and most difficult - aspect of library support for text mining. While we provide all of this wonderful content (databases, ebooks, electronic archives, and more), all of our collections have been built historically for human readers. For computers to "read" our electronic resources means that they have to be able to download (or at least somehow obtain) a very large amount of content, something that might not only overload the servers of the content provider, but almost certainly violates traditional electronic resource contracts, which of course were designed with human readers in mind. Fortunately, excellent historical, literary, and cultural content for many eras and geographies can be found, particularly covering works that are no longer in copyright. Quite a few Open Access (OA) repository resources offer content that resources - digital libraries when the technical process of automatically pulling data off of a web site and creating a structured data set from it. One of the more popular OA resources that provides ready access to digital humanities content is the Making of America (MOA) project, which was a collaboration between the University of Michigan and Cornell University, originally funded by the Andrew W. Mellon Foundation in 1995 (https://quod.lib.umich.edu/m/moagrp/index.html).

In terms of licensed and proprietary content, it has become much more common in recent years for vendors to offer data access, although this is still not a standard contract term for most academic libraries. Moreover, libraries usually license new resources, as well as by seeking addenda on existing resource contracts. The time to ensure access to your collections for text mining purposes is prior to the need, as these licensing processes can take some time, even when the parties readily agree to business terms. For digital collections that are not part of a library, then some reasonable means of data access for purposes of computational research should be included; libraries should avoid investing in content whose use is contractually limited to non-computational readership. For leased collections, the marketplace has not yet fully settled on one model, for the very pragmatic reason that it’s not necessarily reasonable for a text mining project (which may last for many years) to continue to use robust data sets beyond the period of lease access.

Accessible Archives was one of the first commercial vendors of digital archives to offer equal access for text mining as for human readers for perpetual access content. A data mining addendum is included in their standard license agreement. In addition, as Accessible Archives proudly ites on the high quality of its TEI Lite XML and reloaded content at 98% quality or better, their collections more easily suit the needs of novice or occasional textual analysts, which are popular choices among researchers.

High-end data science researchers can handle content in virtually any way that they can get it, although very structured metadata is usually preferred. Technically less sophisticated researchers (which are the vast majority of humanist text miners) generally require fairly well-structured metadata to accomplish their work.

Discovery
Researchers who engage in text mining often have great difficulty in finding accessible data sets. Even libraries that have worked with researchers on this project may not have others who do not yet represent them well. In part, because as a research library community, the frequently thorny issues of discovery have not yet been adequately addressed in any standardized fashion. At present, discovery that data access is available for resources largely happens outside of library catalogs and usually only via lists on library web sites of resources that can be mined by authorized users. This effectively means that researchers who look for resources first within the silo of the type of research they wish to do (text mining) and only then based on the nature (subject, period, genre) of the content itself. Clearly representing the means of accessing data sets also proves challenging, as these are quite diverse, ranging from APIs to mediated vendor requests via librarians to even local library storage on hard drives. In an ideal environment, libraries would readily offer samples of the data set as well so that researchers can make sure that they have the capacity to deal with any particularities of its formatting. Lastly, the provenance and history of the data set (if even known by its owner) may be vital to a given researcher, especially as metadata practices may have changed during the course of its digital production, particularly for resources that were created or revised over a longer period of time. Believe it or not, these are only some of the aspects of a data set that may be crucial to a researcher in finding an appropriate corpus to mine.

Services
Providing appropriate services in support of text mining activities is similarly challenging, and should be very context-driven, reflecting the user needs, mission, priorities, and capacity of the individual library. For high-end researchers, simple access and discovery support may be adequate, as they already have the tools, expertise, and support structures to conduct research on their own. Yet, the majority of our text users (and the fastest growing demographic) fall somewhere in the range between novice and knowledgeable non-expert. Scaling support services to this community on your campus requires knowing that community and recognizing that its needs may change rapidly.

Will your library support text mining tools? Many strong schools (perpetual access) will provide users; one of the most venerable sets of textual analysis tools is Mallett (Machine Learning for Language Toolkit), produced by the University of Massachusetts at Amherst and freely available online at http://mallett.cs.umass.edu/. Many more or less "out-of-the-box" applications satisfy the needs of novice or occasional textual analysts, and easy web searches reveal the most popular tools that other libraries recommend and promote to their users, obviating the need for your library to wholly reinvent this wheel. For more advanced text mining researchers and those who are willing to invest time in learning a more robust tool, the software RStudio, which offers an open source edition, is today’s most popular choice. At the very least, installing basic open source tools and RStudio software on your computer will be helpful to many of your users and begin to communicate at least some level of support for text mining activities.

Will your library provide training for text mining? Many research libraries are now finding that they cannot offer enough instruction sessions to meet the demand for RStudio training. Many routinely offer instruction in web scraping, visualization, and other related tools to users hungry for this content. Of course, whether the library is the appropriate provider of this instruction depends on how your campus is structured, but as a large information need at present, librarians bear a responsibility to at least make sure that it is well addressed at their institution. Perhaps it is best for your library to partner with other units in making sure these issues are scratched, or perhaps even a vended solution fits your institution best. There are a growing number of commercial options that provide on-demand training for users in text and data mining techniques, methods, and tools, and for many schools the cost of licensed training content may scale better than developing, hiring, and supporting staff with the necessary expertise.

At a minimum, consultative needs must be anticipated and addressed, as inevitably users on every campus will seek assistance for text mining activities. Planning ahead with regard to access, discovery, and support services for text mining will at least show that librarians have thought about their roles in providing information services within the realm of computation. These roles should be considered carefully not just within the silo of text mining or digital humanities, but holistically within the larger context of support for digital scholarship and data research of all kinds across the disciplines.
Accessible Archives Responds to Our Customers’ Needs!

Carolina Consortium
Accessible Archives is pleased to join the participating academic and public libraries in the Carolina Consortium! We recently attended the Carolina Consortium Conference and Iris Hanney conducted a successful Premier on Accessible Archives!

Michigan & Pennsylvania County Histories Now Available
Accessible Archives announces the completion of Pennsylvania & Michigan in our landmark American County Histories Series. Accessible Archives is the only publisher that has collected and digitized all of the county histories of the U.S. – all 50 states and the District of Columbia in one database! We offer free MARC records, images and full text of all the books – over a million pages of content!

Expanded Direct Product Links
Accessible Archives has responded to requests from our customers for expanded direct browsing and search links for two of our most popular digital collections – African American Newspapers and American County Histories! Accessible Archives recognizes the value of these expanded links for use in a library’s research guides and libguides.

Achieving Higher Customer Satisfaction Is Our Goal at Accessible Archives

Katherine Brown, Collections Analyst, Auria Library -- “Thank you so much for your help with figuring this out! I really appreciate your prompt responses and dedication to figuring out the problem.”

Elizabeth J. Cronin, Coordinator Information Services, Ocean County Library – “The Military Newspapers of the WWI archive has been great to promote since it includes the Camp Dix paper. The picture of the Camp library is a treasure.”

Barbara Kelly, Director of Libraries, Faulkner University -- “Thank you so much! Thank you for working with us in the way that you have. I have to say, I have never had a vendor work with us so well. We look forward to continuing our patronage with you and marketing the product a bit to our students.”

Angie Thompson, Cataloging Assistant, Liberty University – “I really appreciate your quick response and timely resolution. I deal with a lot of our electronic content vendors when problems arise, and your team’s support is head and shoulders above the rest!”

Upcoming Conference Events
Will you be at the ALA Annual Conference, June 20-25, 2019?
We’d love to visit with you at Booth 3041!
Contact us for an appointment; we have lots to talk about!

Accessible Archives:
is the exclusive sales and marketing agent for Accessible Archives:
Walter E. Washington Convention Center, Washington, D.C.
Accessible Archives, Booth #3041

© 2019 Accessible Archives, Inc.

Accessible ARCHIVES
NEWSLETTER
SUMMER 2019
VOLUME VIII, ISSUE 2

Text Mining in the Humanities: What’s a Library to Do?

So, are your humanists asking you for text mining support yet?
Some Humanities scholars have been engaged with these activities for many, many years now, but more and more seem to be getting involved as interest grows and tools become more user-friendly. Libraries increasingly face questions of how to appropriately support text mining activities. Here is a primer and pragmatic advice for those who are considering just how to do this.

Darby Orcutt is Assistant Head, Collections & Research Strategy, NC State University Libraries, Faculty, University Honors Program, Affiliated Faculty, Center for Innovative Management Studies, Affiliated Faculty, Genetic Engineering & Society Center, and Affiliated Faculty, Leadership in Public Science Cluster, as well as recently served as the Associate Chair of the Faculty of NC State. A national leader in developing models for access to proprietary and use-limited data for content mining and computational research, his current work revolves primarily around research support and engagement for interdisciplinary teams.

First off, what is text mining? Generally, the term is used to refer to a host of computational research practices where a computer “reads” texts at scale and uses algorithms (or artificial intelligence) to quantify or classify texts or their elements. Text mining results can expose linguistic or semantic features, recurrent correlations, and complex patterns across a large corpus. Patterns that a highly observant human scholar might recognize only after a lifetime of reading within their field - or never even notice at all - might be discovered by means of text mining within mere seconds.

Incidence of the word terms “fast” and “fast” in English documents from 1700 to 1900.

Most librarians associate text mining primarily with the digital humanities, but text mining has been an established part of science and social science fields for a very long time. It has perhaps come to the awareness of most academic librarians because humanist scholars have started asking their libraries for help in ways that their more technologically- and quantitatively versed users may not have. In addition, as perhaps disproportionately heavy users of library resources, humanists naturally look to their library as a primary means of discovering information content, accessing it, and for help in using it.

© 2019 Accessible Archives, Inc. is the exclusive sales and marketing agent for Accessible Archives:
Unlimited Priorities LLC®
Iris L. Hanney, President
239-549-2384
iris.hanney@unlimitedpriorities.com
www.unlimitedpriorities.com
Robert Lester, Product Development
Unlimited Priorities LLC
203-527-3739
robert.lester@unlimitedpriorities.com
www.accessible-archives.com

Publisher and editor of Inside the Archives

© 2019 Unlimited Priorities LLC®