Data Mining the Gale Digital Collections

Frequently Asked Questions

GALE
CENGAGE Learning®

What is text and data mining?

Text and data mining, or TDM, is defined by Bernie Reilly, the Center for Research Libraries' President as "automated processing of large amounts of structured digital textual content, for purposes of information retrieval, extraction, interpretation, and analysis."

To Gale, text and data mining represents a new opportunity to lead the way in making content highly useful for scholarly research. Gale is committed to doing whatever possible to facilitate faculty and library connections, take part in new research across the humanities and social science disciplines, and position the library as the center of research.

Why is this important to researchers?

There are many reasons why researchers choose to invest time in TDM. Not only does it represent an almost entirely new front for investigation and analysis, TDM also allows researchers to enrich content through improved indexing, perform a systematic review of literature, create new databases that themselves can be mined, and do research on mining itself for computational linguistics research.

What do new or past archive purchasers need to do to get access to the data?

There is a two-page addendum contract that needs to be signed in order to request the data. The library would need to request this data through their sales rep either at the point of a new purchase or at any time for previously purchased archives.

What is Gale providing, and how is it delivered?

Gale will deliver a hard drive with the XML from each of the resources that the library requests (and for which the corresponding archive purchase has been made). The XML will differ depending on the resource (platform, date created, material type, etc.). Any resource that is cross-searchable on *Gale Artemis: Primary Sources* will have the same data structure. The XML content is the same as the OCR text viewable in the interface as well as accompanying metadata. Please note that Gale will not provide this drive to an individual user at an institution. The library will be responsible for getting their students/faculty access to the content after Gale has delivered the drive.

What does it cost to get the data?

The terabyte drives with data are provided *at the cost of production and delivery*—at this time Gale does not make any additional revenue from this service. Depending on the time required to load the drives and the size of the drive required, this charge can range from approximately $500-$1000 per Gale resource.* Larger or more expansive collections of multiple resources may be more. Gale is developing a list of costs by resource for all that are available at present and expects this list to be ready on December 12, 2014. Once ready, libraries can request this list via their library sales consultants.

*These additional costs apply only to previously acquired resources. Any cost associated with a request for TDM content along with the purchase of a new resource will be rolled into the purchase price of that resource.

How can the data be used, and by whom?

The library will be responsible for managing access to the data. The data can be used by any faculty, student, librarian, or scholar that would have access under the original license agreement signed. The data can be used in myriad ways. Gale does not, at this time, provide any additional tools or server processing space for data analysis outside of those tools already available in the resources themselves. The terms and conditions of use are covered in the licensing addendum.

Do researchers need to let Gale know when they're publishing research based on the data?

No, but Gale would love to know what work is being done.

Will purchases going forward automatically include the TDM hard drive?

No. The TDM hard drive will be sent only if the library requests it and the source institution (the partner with which Gale worked to digitize the content in the first instance) allows it. Another stipulation is that the resource includes significant OCR content. For example, the resource *British Literary Manuscripts Online* offers a compilation of handwritten documents and would not provide a good source of TDM content.

Is there a website where I can find all this information?

Yes, please visit www.gale.cengage.com/TDM to learn more.

Does Gale provide any software to do data mining?

Some light data mining tools are provided through the *Gale Artemis: Primary Sources* interface, such as Term Clusters and Term Frequency Charts, but at this point, Gale does not provide any additional software or functionality. However, this is something that will be investigated going forward.

What's the processing and turnaround time to get the data?

The anticipated turn-around time from request to delivery is three weeks, but this period may be extended slightly due to the size of the request and the number of other requests in the queue.

Are all *Gale Digital Collections* available for TDM?

Unfortunately, no. Some partners (libraries, archives, other content repositories) have either declined to participate or are still working toward an agreement regarding TDM. Gale's hope is that *all* partners will see the academic community's need to access content in this way and will, accordingly, agree to participate in this program.

Whom should I contact if my question is not addressed here?

Ray Abruzzi, Senior Director, New Product Strategy
Ray.Abruzzi@Cengage.com
(917)763-1477

Source Code: 15P-RF0913