eduTDM White Paper

Authors: Bikash Gyawali, Petr Knoth and Nancy Pontika; CORE, Knowledge Media institute, The Open University, UK

Acknowledgement

This work has been conducted with the support of a stakeholder group consisting of publishers, academics, industry and policy organisations' representatives who formed a Working Group (WG). We would like to thank Duncan Campbell, Vicky Gardner and Melissa Harrison for their extensive feedback to the report, WG discussions and their active participation during the consultation. We would like to thank the remaining working group members for their feedback and participation.

Working Group Members

Duncan Campbell – Senior Director, Global Sales Partnership, Wiley

Vicky Gardner – Head of Research Services Development, Taylor and Francis

Melissa Harrison – Head of Production Operations, eLife

Victor Botev – IRIS.AI Co-Founder and CTO

Rachel Bruce – Director Open Science and Research Lifecycle, Jisc

Jacobo Elosua – IRIS.AI Co-Founder and CFO/COO

Victoria Eva – Policy Director, Elsevier

Kathleen Shearer – Executive Director, COAR

Disclaimer

This study does not imply endorsement from the partners' affiliated organisations.

Date of release: 10th February 2020

Section 1: Background

Recent years have witnessed an unparalleled upsurge in the quantities of research articles, with their volume doubling every three years. In the world of knowledge production, researchers worldwide generate over 1.5 million research articles on an annual basis, while over 100 million of such articles had been published as of 2015 (Björk, Roos and Lauri, 2009). While, undoubtedly, these vast amounts of new data and information can offer new insights, give rise to new opportunities for analytics and improved understanding, it is equally undoubted that reading and analysing them is beyond human capacities.

Text and data mining (TDM) is emerging as a powerful tool for harnessing the power of and discovering value in data, by analysing structured and unstructured datasets and content at multiple levels and in many different dimensions in order to discover concepts and entities in the world, patterns they may follow and relations they engage in, and on this basis annotate, index, classify and visualise such content (Knoth and Pontika, 2016).

A study into the Value and Benefits of Text Mining commissioned by Jisc in 2012 (McDonald and Weir, 2012) concluded that text-mining of research outputs offers the potential to provide significant benefits to the economy and society in the form of increased researcher efficiency, by unlocking hidden and developing new knowledge (often by "seeing" links or connections that cannot be glimpsed at a "human readable" scale) and improving the research process and its evidence base. These benefits will result in significant cost savings and productivity gains, innovative new service development, new business models, new medical advancements of research, etc.

Section 2: Motivation for eduTDM

Staff and students affiliated to a university can access and download all the research articles the institution subscribes to, provided that they are logged in to the institution's network. While readers can access research articles their university subscribes to quite easily, it is not possible for text and data miners to machine access research articles their university subscribes to effectively and at scale.

The current amendments and exceptions in the UK Copyright Law have created new opportunities for those who conduct research or study in the UK, who are now able to TDM content provided that they have lawful access to the resource and the research is for non-commercial but research purposes. eduTDM aims to find a pragmatic solution to arrange how the contents of research articles can be delivered to text miners as easily as possible based on the subscription they have. The content served by eduTDM is envisioned to be total data present in research articles (i.e. without fine grained selections of only text, only images, only citations etc.). It shall, however, be delivered in some structured document standard (eg: XML) which text miners are then free to process for retrieving data at their granularity of selection. The aggregation document standard shall, therefore, only act as a thin wrapper to cumulate content as received separately from individual publishers. Likewise, the eduTDM content will only comprise of research articles and not include supplementary resources that may be linked to research articles (datasets, videos etc.)

Section 3: Scope of this work

This report emerged from the first meeting of the eduTDM working group, which took place on the 4th of October 2018.

In that meeting, it was proposed that one or more trusted entities in the UK should be able to provide/broker large scale machine access to all research articles, from across multiple publishers, for TDM purposes for the benefit of staff or students at an institution that has lawful access to this content, thanks to the institution's subscription or due to the content being available through the open access route. We then asked this WG for feedback.

The representatives of this WG are experts in one of the following fields: TDM, publisher systems, industry, and/or making recommendations and creation/implementation of best practice. More specifically, the WG members are:

- Commercial publishers: Taylor and Francis and Wiley
- Open Access publishers: eLife
- Companies performing TDM: IRIS.AI
- Organisations creating TDM corpuses: CORE
- Policy making: COAR and Jisc
- Industry: Knowledge4Innovation

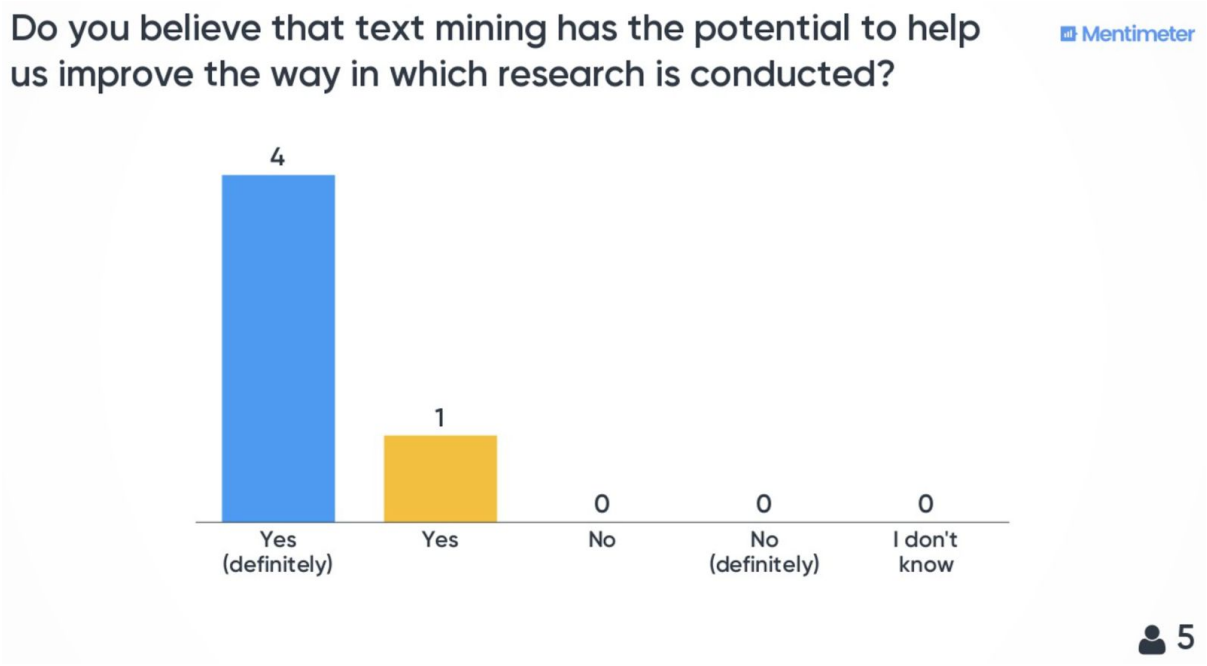
The scope of the initial work was to examine whether the WG participants agree with the eduTDM vision. This report will primarily focus on the technical details required to accomplish an eduTDM network and will not extend to the organisational and governance challenges.

Section 4: Results from the first eduTDM meeting

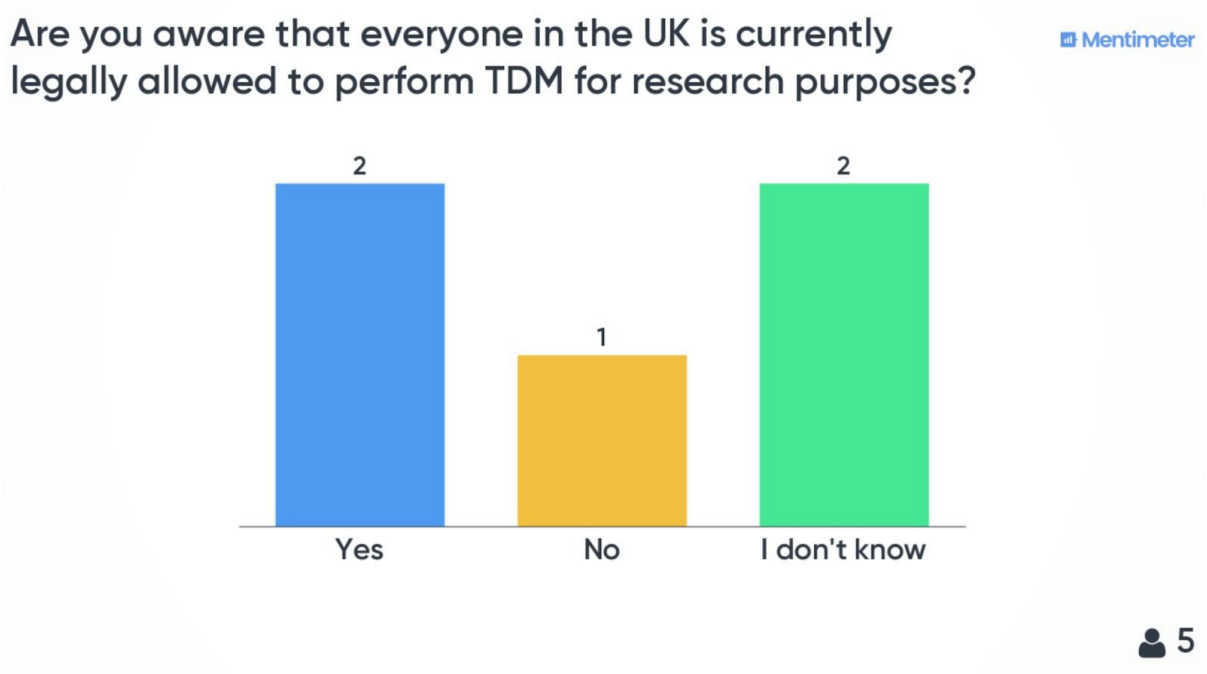
During the first eduTDM meeting, the WG participants were asked to respond to and comment on three sets of questions. The first two sets were closed ended and attempted to examine whether the WG participants were aware of the UK TDM Copyright Exception and realise whether the WG participants were in favour or against the eduTDM idea as it was proposed. The third set was open ended and asked the participants to comment on the technical and organisational challenges.

The closed ended questions asked were:

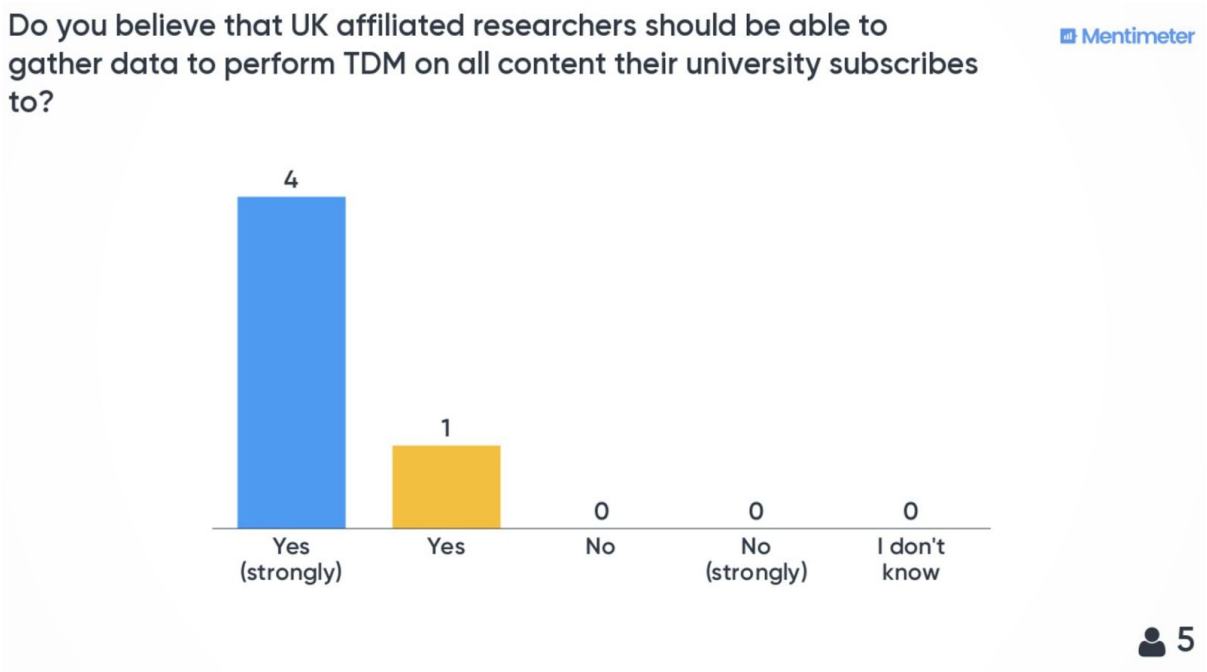
1. Do you believe that text mining has the potential to help us improve the way in which research is conducted?



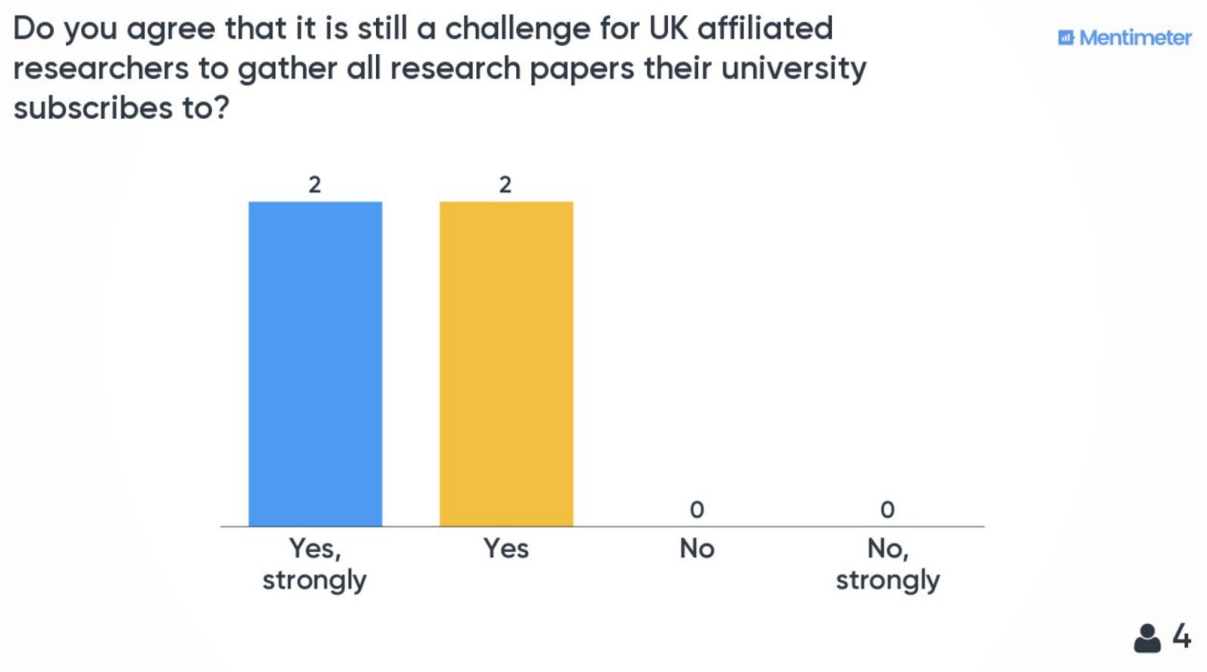
2. Are you aware that everyone in the UK is currently allowed to perform TDM for research purposes?



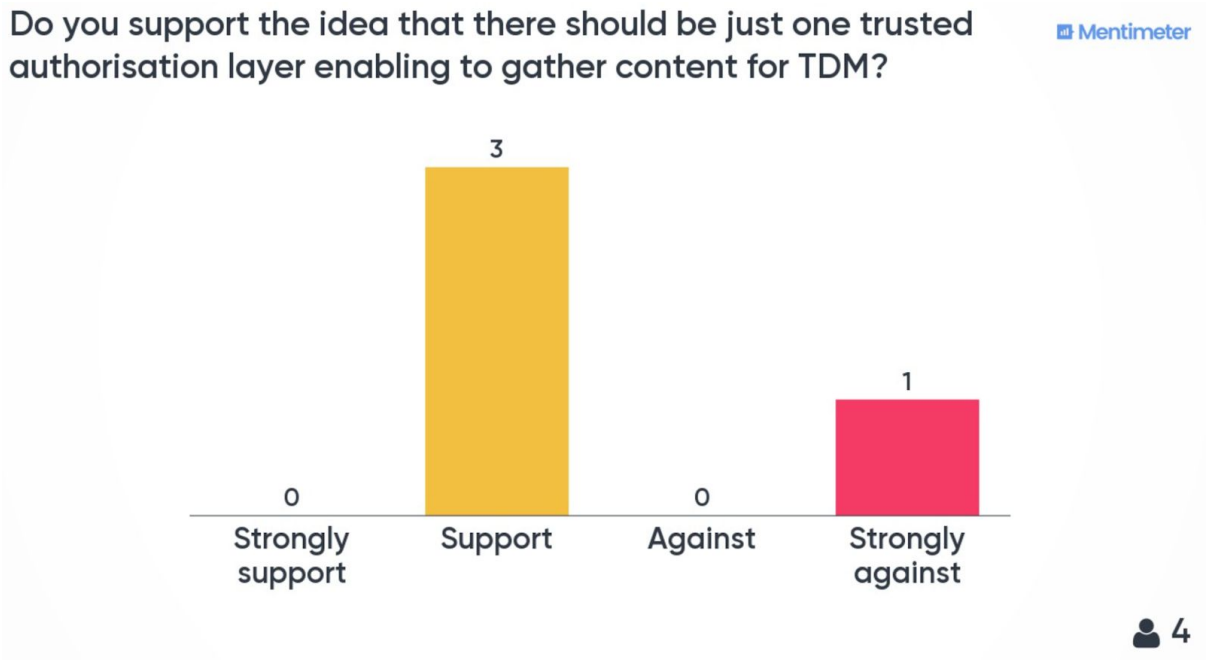
3. Do you believe that UK affiliated researchers should be able to gather data to perform TDM on all content their university subscribes to?



4. Do you agree that it is still a challenge for UK affiliated researchers to gather all research papers their university subscribes to?



5. Do you support the idea that there should be just one trusted authorisation layer enabling to gather content for TDM?



The open ended questions asked were:

A. What technical challenges do you foresee in establishing TDM?
B. What organisational challenges do you foresee in establishing TDM?
C. What is the greatest issue to overcome in establishing TDM?

With regards to the results, all WG participants mentioned that they are aware of the many possibilities and benefits emerging from TDM (Q1), while not all of them were aware of the recent Copyright Exception in the UK, which is limited to research only and not commercial use (Q2). For this question there was a discussion about the word 'lawful', as this word was missing from the question given, but is mentioned in the updated legislation prohibiting the use of content to those who do not have a subscription. In addition, the WG participants responded positively and strongly to the third question (Q3).

All WG members agreed that there are challenges for the universities in collecting the lawful content for TDM (Q4), and there were concerns about the idea of "one trusted authorisation layer" that enables the creation of a corpus for TDM purposes (Q5). A clarification was then provided, highlighting that the goal was to implement one trusted protocol or process, rather than having only one entity that acts in this area. In the open ended questions the WG participants had the opportunity to provide input by adding free text. The primary focus of these questions was the challenges in putting the eduTDM idea in place and brainstorming on how this process could be implemented. The responses with regards to the technical challenges (QA) mentioned a paper's format, i.e. PDF, XML, HTML, the lack of common standards for metadata, data and APIs and the need for an agreement on them, managing subscriptions, and lack of a common registry of API subscriptions at an institutional level.

The responses on organisational challenges (QB) relate to the incentives that the publishing industry should consider to adopt the service and agree on common standards, burden sharing, e.g. infrastructural costs, perceived benefits, common legal frameworks, gaining traction. When the WG

participants were asked to provide their comments on a single and perhaps the greatest challenge (QC), their responses focused more on the organisational challenges than the technical. This indicates that even though a technical solution is possible, it is the organisational challenges that would have a negative impact on the progress of this initiative.

To conclude, the WG participants are in favour of eduTDM, understand its capabilities and benefits, and are keen on the idea that researchers affiliated with a UK university should take advantage of the recent amendment in the UK copyright law. They all agree that currently researchers affiliated with a university face challenges in the lawful collection of a large corpus of data for TDM purposes and one or more trusted entities could possibly provide a solution. Nonetheless, there are many technical challenges, such as standards and formatting, and organisational, such as infrastructural costs and gaining traction, with the most difficult appearing to be the organisational ones.

It was discussed and agreed that the CORE Team, which has proposed the eduTDM concept and has put the WG together to discuss this topic, will now focus on drafting a proposal for addressing the technical challenges, as presented in the below section, demonstrating how the eduTDM vision can be technically realised.

Section 5: How can this work - technical concept

We envision eduTDM to serve user requests by aggregating content from multiple publishers. There are two main tasks that underpin this service. The first is the task of delegating user requests to publishers endpoint. This is necessary to ensure that only relevant, authorised and updated content from publishers are generated for each user request **directly by respective publishers**. The second task which occurs **locally on the eduTDM's end** is to aggregate all such content received from publishers and serve the aggregated content as end result to user query. Before delving deeper into the details of these tasks (in sections below), we begin by identifying different stakeholders that must become involved in any of those tasks and the functionalities they desire. At present, we can identify at-least 3 such entities :

i) Individual users (**Researchers**[1]) interested in scientific text/data mining.

ii) **Universities** holding subscription to one or more publishers.

iii) **Publishers** that publish electronic copies of research articles such as journals, conference proceedings, etc. Such content can be open access content or subscription only content.

Having identified the various entities involved, we can now consider the functional requirements for the eduTDM service. We can outline these below in the form of user stories.

Functionalities desired by Researchers :

- As a researcher, I want to use a single endpoint so that I can download all content to which my university subscribes to or that is open access using a single API.

[1] Throughout this manuscript, the term researcher refers to researchers performing large scale text/data mining on scientific publications.

- As a researcher, I want to be able to rely on my university authentication mechanism to access content from across publishers.
- As a researcher, I want to be made aware of content existing on publishers' platform which matches my requests but is unavailable for access due to limitations in my university's subscription to the publisher.
- As a researcher, I want to have all content delivered in a uniform format, such as PDF and/or XML as produced by the publisher, so that I can text/data mine using a single (same) pipeline of operations.
- As a researcher, I want to be able to receive incremental downloads so that I can easily update my collection as soon as new content becomes available.

Functionalities desired by Universities :

- As a university, I want to ensure that all my users can text and data mine all subscription based or open access content from all the publishers I subscribe to.
- As a university, I want to ensure that all my users can text and data mine all open access content.
- As a university, I want to have information on usage statistics of eduTDM by my users.

Functionalities desired by Publishers :

- As a publisher, I want to have very minimal or no changes to my existing model of content delivery so that adaptation to eduTDM becomes practical.
- As a publisher, I want to have a secure communication channel so that my content is guaranteed to be delivered to users with valid access rights only.
- As a publisher, I want to have service guarantees so that reliability and efficiency can be ensured.
- As a publisher, I want to have the ability to monitor usage of my contents delivered through eduTDM.
- As a publisher, I want to inform my users whenever they are missing some of my contents due to limited authorisation (subscription) rights they bear.

We can now identify the requirements which must be fulfilled by the eduTDM service to meet the functionalities just discussed.

Requirements related to Researchers :

Functional Requirements :

- The service shall provide users with one point of access to get content from across multiple publishers.
- The service shall ensure that only valid users shall be able to issue text/data mining requests.
- The service shall aggregate content from all the publishers to which the user is authorised and deliver it.
- The service shall deliver its content in a universal format.

- The service shall provide users with means to request incremental content such as fetching new content since their last download date.
- The service shall meet user specified constraints(filters) while looking for matching content to deliver.

Non-Functional Requirements :

- The service shall be able to handle (TBD) number of users per minute.
- The service shall implement (TBD) encryption mechanism to secure content delivery channel to users.
- The service shall inform users when they are missing access to any content from any publishers because of limitations in their subscription rights.

Requirements related to Universities :

Non-Functional Requirements :

- Publishers shall allow universities to use the eduTDM service.
- The service shall not interfere with other activities (services) of universities.
- The service shall be able to provide usage summary to universities on a quarterly/monthly/daily/online (TBD) basis.

Requirements related to Publishers :

Functional Requirements :

- The service shall guarantee to act in compliance with licenses to the provided content while delivering it to end users.
- The service shall be able to ask for incremental updates of contents from publishers.

Non-Functional Requirements :

- The service shall be able to deliver large amounts of content for TDM without posing a high load on the existing infrastructure of publishers, i.e. protecting existing content delivery models used by human users.
- The service shall be able to provide usage statistics to publishers regarding content accessed from them.

Section 6: Proposed Design

Our proposal for the eduTDM architecture is motivated by the widely popular solution for internet access across universities -- the eduroam service. eduroam provides its users a common platform to access the internet from across a network of universities. Similarly, we would like to propose the eduTDM framework as a one-point access platform for our users to access content from multiple different publishers for text and data mining.

Fig 1 below depicts the current state of practice which researchers adopt in gathering content from multiple publishers. A user interested in text mining of content from different publishers typically issues separate API calls to each of the publishers, as shown in the figure.

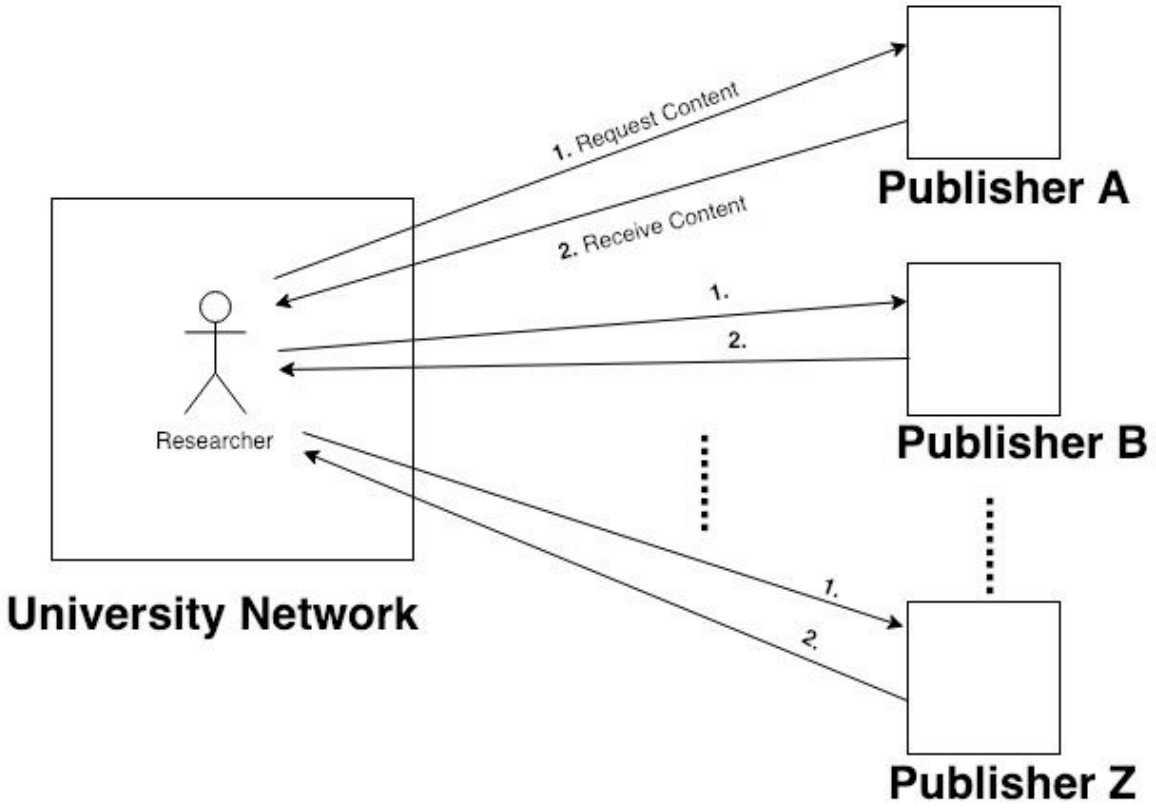


Figure 1 : Traditional model of content mining from multiple publishers.

There are, however, several issues that make this practice a cumbersome experience. We outline them below.

A. **Abundance of Publishers** : There are several hundred publishers in existence today. Many of them expose their own API endpoints. This imposes two main problems to researchers. First, they need to be individually aware of existing API endpoints and in some cases the content is not available via any API. Second, they need to write and maintain multiple copies of similar queries to issue to each publisher separately.

B. **Heterogeneous Standards** : Different publishers implement different protocols for users authentication and content delivery. Often, this implies that researchers create an account on each publishers' platform and ensure that correct authentication tokens (e.g. API Keys) are used when issuing requests to respective publishers. Further, publishers can adopt different standards for content delivery. Therefore, a researcher interested in getting the same type of content, for example on global warming, from multiple publishers will have to write and issue API calls to each publisher in different formats, using different accounts and will receive responses also in a variety of formats.

C. **Workflow Disruption** : From the users' perspective, writing multiple API calls deviates focus from their main line of work. Their main goal of TDM would be better facilitated if there were a direct means of obtaining aggregated content from publishers world-wide.

To summarize, with the existing practice, users are individually responsible for maintaining their system and need to adapt for different publishers and the heterogeneous standards they expose. This is prone

to manual error, is limited by domain knowledge of individual researchers and doesn't automatically scale up to accommodate changes/updates introduced by publishers in the future.

eduTDM aims to provide a better alternative for machine access for TDM of content from multiple publishers. Figure 2 below outlines the components of this architecture.

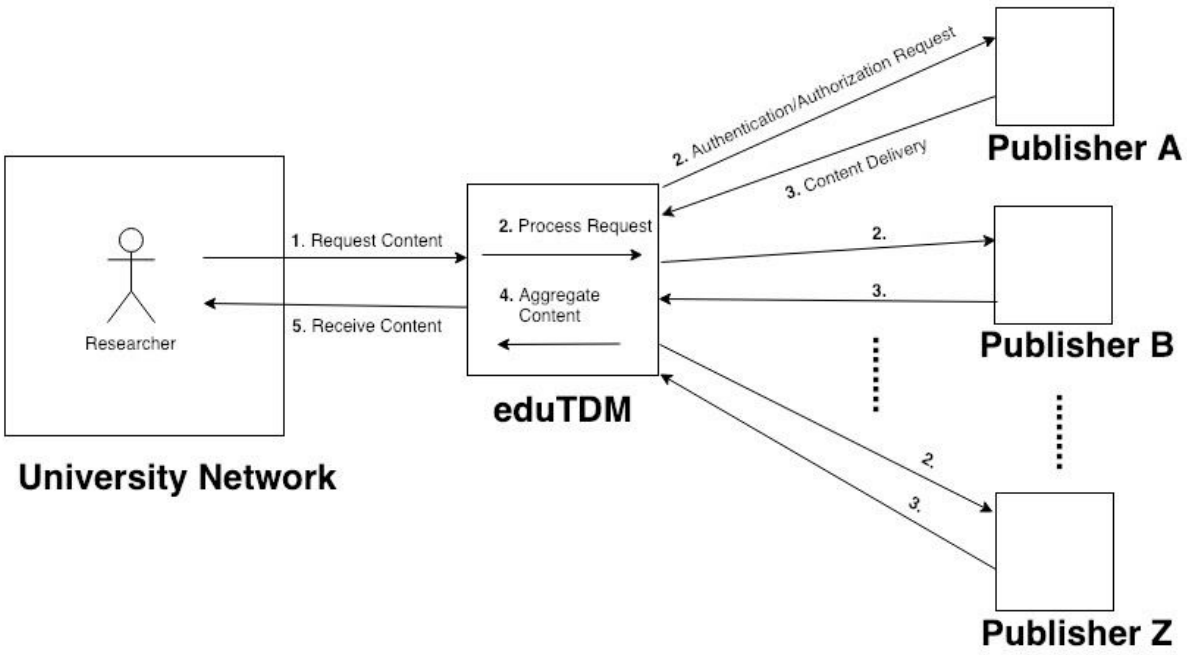


Figure 2 : eduTDM vision for scalable content mining from multiple publishers.

In the eduTDM model, users can issue a **single** API call to eduTDM to obtain **all** documents matching their need. This will involve a workflow comprising of a sequential series of steps as follows:

1. User issues a "request for content" call to eduTDM.
2. On behalf of the user, eduTDM issues separate content request calls to each of the publishers.
3. Each publisher independently processes its incoming request and responds to eduTDM with the requested content.
4. eduTDM aggregates and harmonises the content from all the responses received.
5. eduTDM responds to the user with the aggregated content.

We can now think about technical solutions that can be used to implement the workflow. For each step of the workflow outlined above, we present our proposal solution as follows (preserving the sequential order of steps in the workflow):

STEP 1: Content Request

User requests must be supported by means of an API exposed by the eduTDM service. Users must first register to this API to get an API key which will be used in further communications as authentication token. This API shall implement the following features :

A. Only accept traffic originating from networks of subscribing universities. eduTDM must gather the information on universities' IP address pools when universities first subscribe to use this service.
B. For each incoming request, extract the information about its originating network (i.e. its IP address). This information is required to fulfill step 2 in the workflow (discussed below).
C. Allow users to specify filter criteria, e.g. look for partially or completely matching documents on a given topic, get only incremental updates since the last execution, exclude content from certain publishers, etc.

STEP 2: Processing Requests

For each incoming request, eduTDM must carry out the following tasks:

A. **Validate the request:** This shall be based on verifying that the API key used for the query belongs to some registered user of the eduTDM service.
B. **Create calls to multiple publishers with URL redirection:** At this phase, eduTDM needs to issue content request calls to multiple publishers. This task can be divided into several subtasks:
    i. Identify content request fields and filters specified in the incoming call.
    ii. Create separate copies of query to each publisher: All such copies must contain the document request criteria originally specified by the user but also be specific to meet the requirements of specific publishers. eduTDM must adopt the data request/exchange standards of each publisher to identify the means and format of queries they support for document delivery. As of present, many publishers specify their own API endpoint for text/data mining purposes. We are aware that publishers may serve their APIs on publicly accessible endpoints or as internal links to local index of their content within a university/organization. In the former case, the endpoints could be hosted on the publisher's proprietary platform or on shared publishing platform (eg. Atypon, Highwire). In the latter case, publishers shall provide appropriate interface for eduTDM to communicate with. This will allow eduTDM to make requests to the local index of publishers' content. In both instances, carrying out this subtask implies that eduTDM shall follow the publisher specific guidelines and draft a query matching such specification.
    iii. Enable URL redirection**:** The next step is to issue those calls from eduTDM to respective publishers. Most publishers already have existing mechanisms to authorise requests based on IP addresses of their subscribing universities. eduTDM can exist independently of any university network and have its own range of IP addresses. Hence, we need a mechanism by which calls forwarded by eduTDM on behalf of a user are treated the same as originating from the user's university. External services such as EZproxy, OpenAthens and Shibboleth

can be used in this task. It should be noted that application scenario akin to our needs already exists within existing infrastructure, although in a different setting. Consider, for example, a user **X** affiliated to a university **U**. The university **U**, in turn, has subscriptions to publisher **Z**. When **X** is within the network premises of **U**, she can freely browse the publications in **Z** but when **X** tries the same from the outside world (e.g. a restaurant), the content from **Z** is inaccessible to her. In that case, **X** can make use of an external web proxy server (such as EZproxy), which essentially redirects her request through **U**'s network so that the publisher authorises it as a valid request. eduTDM would benefit from having such url redirection facility to authenticate with publishers. Other third party solutions proposing alternatives to IP address based authentication are also available; such as the OpenAthens platform or the RA21 protocol. Perhaps a good strategy here would be to discuss our use cases with specific publishers and identify the best solution to adopt.

URL redirection must be enabled for every incoming request to eduTDM because different users have different access rights to different publishers and eduTDM must ensure that users will only get those contents for which they have legal rights to access.

Both the validation and URL redirection services contribute to the required set of services which eduTDM must implement.

STEP 3: Content Delivery

Thanks to the URL redirection mechanism discussed above, publishers receiving requests from eduTDM shall treat them as if they originated from the host university of the original user. To process an incoming request and deliver reponse, all publishers must individually implement the following services :

- Authentication: Authentication is the task of validating incoming requests -- determining whether an incoming request should be processed or not depending upon the credentials it contains. Publishers shall individually perform this task for any requests forwarded by eduTDM and make use of the ip address for authentication.
- Authorization: Authorization is the task of determining what content should be accessible to serve an incoming request. It is the responsibility of individual publishers to determine the content access rights for each request forwarded by eduTDM and they should make use of university subscription information to do so. In case a publisher contains some content (relevant to the user's request) which, however, can not be delivered to the user because of his/her limited subscription rights, the authorization service can (optionally) inform users of such content.
- Searching: Of all the content that is accessible, only those that are relevant should be identified to make up a response for an incoming request. Publishers must, therefore, implement some searching service which determines what contents match to generate response and deliver them.

All these services belong to the set of required services that publishers must provide for eduTDM to function and it is understood that this infrastructure is already in place. Many different publishers, in fact, already expose web API endpoints for serving text/data mining queries and eduDTM shall build upon these resources. Table 1 below lists out some major publishers along with the nature of the APIs they support[2]. These publishers may host their web APIs within or outside their proprietary systems (e.g via 3rd party hosting systems such as Atypon, Highwire etc). For eduTDM's purposes, the communication model remains the same as long as the requirements described above are fulfilled. Some publishers, however, may not host a web API but offer alternative schemes of content delivery; such as by providing an index of their contents locally to eduTDM. In such cases, the eduTDM workflow remains the same except that content delivery requests are issued directly over the local index rather than as web API requests and the publisher shall provide appropriate interface for eduTDM to query over its local index.

Publisher	Web API Endpoint Available	Response Format	Content Served
Elsevier	Yes	XML/JSON	Full text + Metadata
Springer	Yes	XML/JSON	Full text + Metadata
Wiley	No	--	--
Taylor & Francis	No	--	--
axXiv	Yes	Atom	Metadata
IEEE	Yes	XML/JSON	Metadata

STEP 4: Aggregation

One or more publishers may respond with their content. It is the task of eduTDM to gather contents received from all publishers and deliver the aggregated content as output to the user's API call. We call this as an aggregation service and it belongs to the set of required services that eduTDM must implement. The aggregated response from eduTDM can be in one of the several standard document formats (e.g. XML, JSON). In the simplest case, it shall be a structured concatenation of responses received from each publisher as they are and without any processing by the eduTDM. Appropriate markup tags shall be used in the aggregated eduTDM response to identify the source publishers of individual responses, respectively. It should be noted that some publishers may only have metadata information for some or all of their content (as opposed to full text). eduTDM shall, therefore, have the flexibility to accommodate such variations in the aggregated content.

[2] A detailed list can be obtained from
https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/

Additionally, we can think of some **optional** services which eduTDM may implement. These services not necessary for the core functioning of eduTDM as described above but when present shall help to improve eduTDM's response time and/or quality. These include :

- Caching: To reduce workload on publisher's endpoint and to improve user response time, eduTDM could be designed such that it doesn't issue exact same requests to publishers repeatedly. The idea here is that eduTDM could keep a local copy of contents which were obtained from publishers in the course of serving past user queries. The cache should reflect the most popular/frequent results delivered by eduTDM in the course of serving responses over a set interval of time after which it shall be refreshed. The cache is, therefore, neither a mirror copy of any publishers' content nor solely comprised of contents only from a specific publisher at any given time. eduTDM shall respect publishers' option to deny caching of their content. Moreover, any content from the eduTDM cache shall be used in generating end response only after such content has been authorised for the incoming user query by the publisher owning that content. In other words, the cache shall only be used to limit flow of duplicate traffic from publishers end point to eduTDM without sacrificing the authentication/authorisation checks from publishers. Publishers can therefore benefit by only delivering differential content (pertaining to the input request) which doesn't already exist in eduTDM cache at that instant.
- Searching: In addition to the general use case of delivering responses by aggregating research articles from different publishers, eduTDM could expose a search functionality on top of the aggregated content. This implies eduTDM conducting text mining activities (on behalf of users) over the aggregated content built to serve input user query. For this, eduTDM can apply standard parsers to extract relevant content from individual publishers' responses and aggregate them to generate the final response to deliver. This could be useful, for example, to retrieve summary data of the aggregated content. Another use case could involve defining constraints which span across research outputs (obtained from multiple different publishers) present in the aggregation. For example, finding most recent research articles, grouping research articles based on their field of study, identifying key phrases in full texts, etc.
- Usage Monitoring : eduTDM can maintain a local database tracking the number and type of queries served in a given period of time. This usage statistics can then be shared with all the stakeholders concerned.

STEP 5

The user receives the aggregated content and proceeds to text/data mining.

In light of this discussion, we can now model the eduTDM architecture in terms of services that are needed for a successful workflow. This is shown in Figure 3 below.

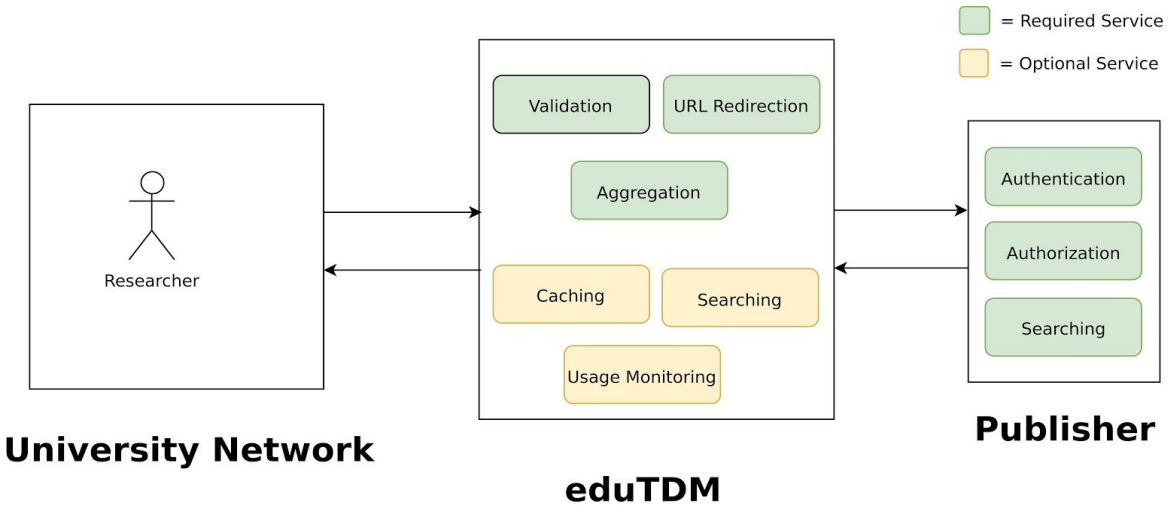


Figure 3 : Services needed for eduTDM implementation.

The key benefits eduTDM offers to researchers over their conventional approach to large scale data/text mining are:

1) **Universal point of access**: Different users from different universities can use this service.
2) **Ease of use**: A single API call to gather content from multiple publishers.
3) **Simplicity**: Users don't need to adapt to multiple publishers' specifications.
4) **Consistency**: The manner in which users interact to get documents will remain consistent across publishers and document domains.
5) **Single point of update/change**: To adapt to future changes in publishers' specification, central changes in eduTDM is sufficient; no more custom updates required on every user's end.

Section 7: Direct comparisons with existing services:

Having discussed our model and the TDM workflow it supports, here we want to provide a direct lookup table for easy comparison of eduTDM with other services in use today. Specifically, we want to compare against a) Publisher specific APIs, b) Crossref[3] and c) Digital Scholar WorkBench[4]. Of the first category, many are in existence today while Crossref and Digital Scholar WorkBench provide TDM support for content across publishers.

Comparison Criteria	**Publisher's API**	**Crossref**	**Digital Scholar Workbench**	**eduTDM**
Content Across Publishers	No	Yes	Yes	Yes

[3] https://www.crossref.org/
[4] https://tdm-pilot.org/

API Output	Metadata + Full Text	Metadata + Link to Full Text	No API but bulk data download possible. No Full Text.	Metadata + Full Text
Accessing subscription based content	Separate authentication tokens (API Keys) per publisher.	Two-step process -- 1) get fulltext link from CrossRef 2) issue request per publisher	Probably not -- it was possible to download the result dataset without any authentication.	Automatic identification of access rights across publishers
TDM Code Maintainability	Users need to update their code to changes in publishers' end.	Users need to update their code to changes in publishers' API endpoint -- Crossref only provides link to fullText.	Easy maintainability and good support for data analysis.	eduTDM shall maintain its code to adapt to publishers' update -- end users shall not be affected by such changes.
Query support	Mainly metadata matching queries.	Mainly metadata matching queries.	Limited to word frequencies (from text fragments of publications which match the input search term) and metadata processing.	Supports queries for matching metadata values as well as analysis of the aggregated content.

To summarise, we argue for a new model to support researchers in text/data mining content from multiple publishers. We have proposed our eduTDM model as a better alternative to the existing practice and outlined key benefits it will bring to end users. We have proposed a technical workflow to implement eduTDM as a Software as a Service (SoA). We invite all WG members to study this proposal architecture and contribute by providing suggestions, comments or further analysis of the workflow.

References

Björk, B.C., Roos, A. and Lauri, M. (2009). Scientific journal publishing: yearly volume and open access availability. *Information Research*, 14(1). Retrieved from https://helda.helsinki.fi/bitstream/handle/10227/615/bjork_roos_lauri.pdf

Knoth, P. and Pontika, N. (2016). Text and Data Mining Taxonomy. Retrieved from https://www.fosteropenscience.eu/taxonomy/term/191

McDonald, D. and Weir, G. (2012). *The value and benefits of text and data mining*. London: Jisc. Retrieved from https://www.jisc.ac.uk/sites/default/files/value-text-mining.pdf