IFLA WLIC 2019 Athens

Submitted on: 24.07.2019

Yewno: Transforming Data into Information, Transforming Information into Knowledge

Philip Schreur
Technical Services, Stanford University, Stanford, USA.
E-mail address: pschreur@stanford.edu

Abstract:

Libraries have long been concerned with the transition to the Semantic Web. Studies have shown that many patrons begin their search for library materials on the Web as opposed to a local discovery layer, making the representation of library metadata on the Semantic Web essential for a library's survival. Most often this is accomplished through the transformation and serialization of a library's bibliographical metadata encoded in Machine Readable Cataloguing, or MARC, to linked data in a commonly used ontology such as schema.org or BIBFRAME. But this linked data is derived from expensive, handcrafted bibliographic surrogates for a shrinking percentage of resources in a library's collection. As full-text digital resources and datasets begin to dominate library collections, this handcrafted approach to bibliographic metadata creation cannot scale. Yewno, through such offerings as Unearth, complements traditional library discovery by providing structured access to these burgeoning collections through the use of Artificial Intelligence.

Yewno begins by extracting entities such as names, places, and dates from digital text. In addition, Yewno extracts concepts from the entire data store and relates them to each other in large graphical structures. Numerous features such as journey mapping, knowledge map layering, and concept expansion allow the user to explore the unstructured data making use of concepts extracted directly from the data. The presentation will include a live demo of Yewno's capabilities (with canned back-up in case of poor connectivity).

The world of discovery and access is in a time of impatient transformation. Through the conversion of metadata surrogates (cataloguing) to linked data, libraries can represent their traditional holdings on the Web. But in order to provide some form of controlled access to unstructured data, libraries must reach beyond traditional cataloguing techniques to new tools such as artificial intelligence to provide consistent access to a growing world of full-text resources.

Keywords: Semantic Web, linked data, BIBFRAME, artificial intelligence, Yewno.

Libraries worldwide have relied upon Machine-Readable Cataloging (MARC)-based systems for the communication, storage, and expression of the majority of their bibliographic data. MARC [1], however, is an early communication format developed in the 1960s to enable

machine manipulation of the bibliographic data previously recorded on catalog cards. Connections between various data elements within a single catalog record, such as between a subject heading and a specific work or a performer and piece performed, are not easily and therefore not usually expressed as it is assumed that a human being will be examining the record as a whole and making the associations between the elements for themselves.

MARC itself was a great achievement, eliminating libraries dependence on card catalogs and moving them into a much needed online environment. It allowed for the development of the Integrated Library System, or ILS, and great economy in the acquisition, cataloging, and discovery of library resources. But as libraries transition their metadata to a linked-data based architecture that derives its power from extensive machine linking of individual data elements, the former reliance on human interpretation at the record level to make correct associations between individual data elements becomes a critical issue. And although MARC metadata can be converted to linked data, many human-inferred relationships are left unexpressed in the new environment. It is functional, but incomplete.

But assuming that this transition from MARC to the semantic web is the correct decision for libraries, why linked data? First, it is apparent that library patrons have preferred searching for information on the Web for quite some time. By integrating library data into the Web in a semantic way, our patrons can find well-formed library data there as well as in library catalogs. By taking advantage of the semantic web, library patrons can directly benefit from other important data sources on the Web. A third advantage is that the Web is an international environment. By shifting to linked data, libraries worldwide can take advantage of the bibliographic and authoritative data many national libraries create and make available now as linked data. And last, the Web is a continually evolving environment. Without a doubt, linked data will evolve into some other standard with time. But in order to move along with this evolution, libraries will need to make that first important step in the transition to a Web environment.

Linked data itself, however, will create its own impacts on Technical Service's processing. Over the years, we have honed our workflows to be ruthlessly efficient. Current processing is run with the brutal efficiency of a Tesla factory with each element of the workflow shaped carefully with very little tolerance for variation. And as additional requirements are added, whether it be reduced staffing or the addition of the digital library, they are added to an already overburdened workflow. We simply can't afford to be less efficient than we already are. But as Technical Services moves from an insular service model confined by the MARC formats to a broader, community-based model rooted in the Web, old processing models begin to break down. Two key areas to examine are the need to convert vast quantities of legacy data to linked data and, in addition, how we support our controlled headings in a new, open Web environment.

Legacy data in the MARC formats will be with us for decades to come. The transition to linked data will be similar to the transition from the card catalogues to the online catalogue. Many libraries participated in grants for the conversion of their card catalogue with all the attendant fears of a new technology, fears of loss of data, fears of loss of jobs. As with the transition from catalog cards to online data, the transition from online data to linked data will also be a lengthy one. Even if we wish to transform our metadata backbone to linked data, many vendors still will wish to supply data in the MARC formats and many library partners will continue to work in MARC for years to come. For instance, the Bibliothèque nationale de France has remained committed to Intermarc for the near future and sees no need for an immediate transition to linked data. Instead, they will rely on conversion to linked data for metadata communication

outside of the BnF but will still create data as Intermarc for internal use. Many libraries may choose to follow this same option and only convert data when needed in this period of transition.

To make matters more complex, MARC isn't only used for metadata creation and exchange. Much of a library's basic functionality is driven by services that make use of the intricacies of MARC's fields, subfields and indicators. Everything from complex internal reporting to discovery is driven through their use in common system architectures. Some functions, such as discovery, can very well benefit from a linked data approach, but others, such as payments or authentication, may function much better making use of MARC data and a relational database.

The transition to linked data will not be complete with the conversion of our legacy data. Many libraries subscribe to metadata enhancement services such as the Nielsen Bookdata service that supplies addition tables of contents or book reviews to resources we have already cataloged. Metadata may also be updated or corrected over time. Headings for people or organizations may merge or split as more information is received. Metadata is a living thing, and all these changes must be captured and reflected in the linked data that we have already published after we convert it. Issues such as what is an update, what is a correction, and what is a deletion, must be very carefully worked out in a linked-data context, especially in a shared, communal environment.

And if we are to convert our data, we must also consider the particular ontology into which we wish to convert our MARC data. By its nature, the Web is distributed and variable. By their nature, libraries prefer to be more consistent for the easy exchange of data. There is a natural tension between the two and choice of which ontology to use is part of that tension. Currently, an ontology initially created by the Library of Congress called BIBFRAME [2] appears to best capture the data we have recorded in MARC but there are other common standards as well such as schema.org [3], Dublin Core [4], or CIDOC-CRM [5]. And even if libraries decide to settle on BIBFRAME as a common ontology, will BIBFRAME become as variable as MARC has been with MARC21, Canadian MARC, Chinese MARC, Intermarc, etc.? As we transition to a Web environment that is variable by nature, how much standardization can we, or should we bring?

Support for controlled headings, or what libraries have called authorities, is another standardization technique linked data brings into question. Traditionally, libraries have used authorities to support controlled headings in their MARC cataloguing but authority record creation is complex and restricted to a small subset of librarians. And these authorities do far more than identify an entity like a person, they provide a preferred form of name, cross references, links to earlier and later headings in the case of complex corporate bodies; they are a wealth of information. As libraries transform this aspect of their workflows, they will need to struggle with a number of important issues.

The first is that a traditional library authority cannot be used as an identifier for an entity such as an author in a linked data context. Authorities are descriptions of people, of places, of events. And since they represent a description, they cannot identify the person themselves as a physical being. And since only people themselves can take an action, such as fulfil a role as an author, you need a different identifier to identify the person themselves as a Real World Object, or RWO.

A second issue is the high-level training required to create a traditional library authority. Authority record creation is often limited to professional staff that have gone through extensive training in the Name Authority Cooperative, or NACO [6]. Their creation is costly and so many entities are not support by a traditional library authority. In a linked data context, however, every entity will need an identifier and these identifiers could be created by anyone in the world. What will be the overlap between the world of authority creation and identifier creation and management? Does the libraries pre-emptive approach to authority record creation have a parallel on the open Web? Current cataloguing rules, Resource Description and Access (RDA) [7], or cooperative programs such as the Program for Cooperative Cataloging (PCC) [8] or CONSER [9] require authorities in support of controlled access points in our cataloguing workflows. Will identifiers alone ever be an acceptable substitute for an authority? Will identifiers created with new partners such as Wikidata [10] ever become an acceptable substitute for traditional programs such as NACO?

And yet, even with the difficulties of the transition of libraries' metadata to linked data, even with the need to rethink key workflows such as cataloguing and authorities, linked data will bring tremendous advantages by finally harmonizing library metadata with the semantic web. In many ways, the conceptual transition from controlled access points to identifiers is not a difficult stretch. Both are meant to link, both can be difficult to create and reconcile. And both, at least considering current common infrastructure design, require some degree of human creation and maintenance.

Yewno [11], however, takes a different approach to resource access and discovery. According to Yewno's homepage:

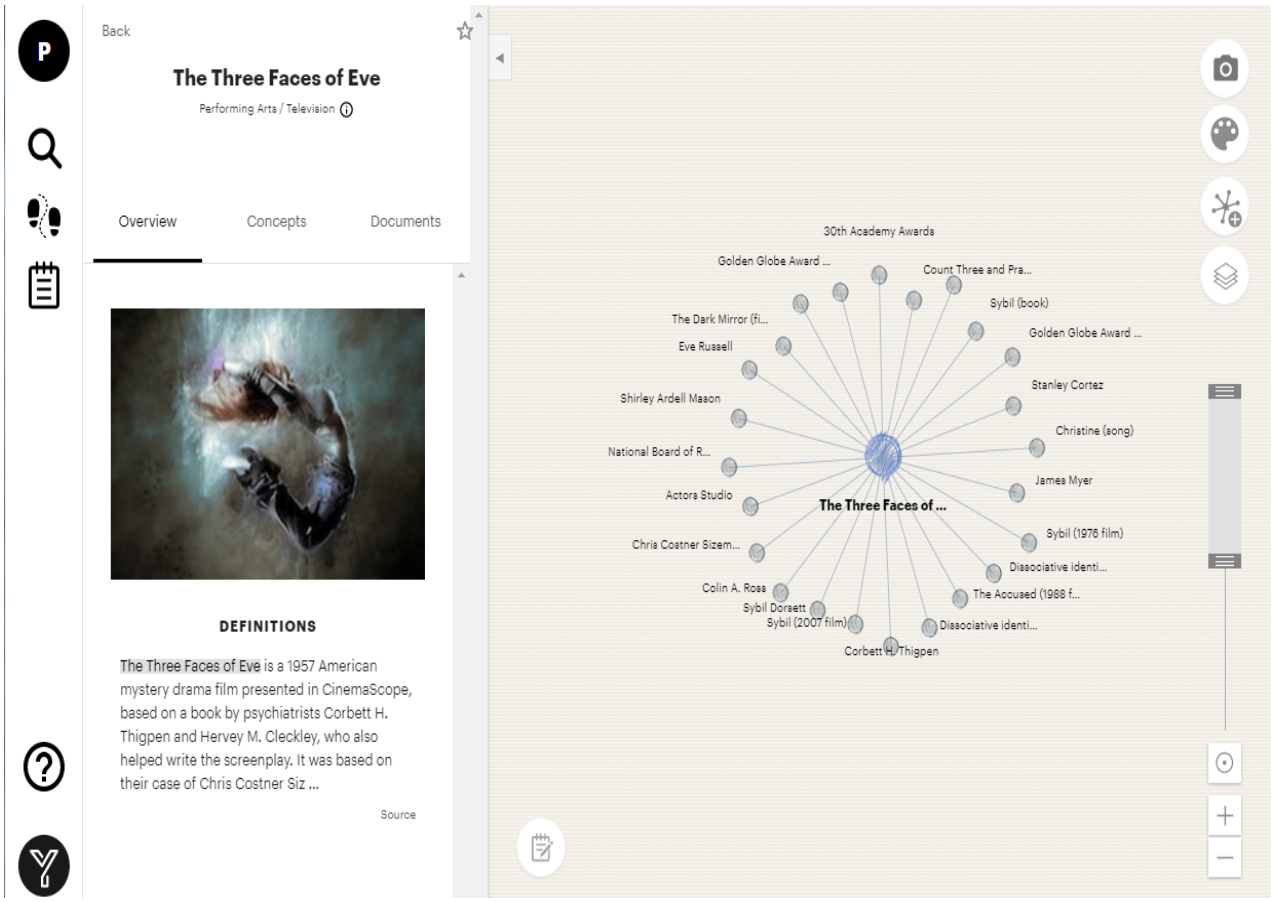
> using machine-learning and computational linguistics, Yewno's unique technology analyzes high-quality content to extract concepts, and discern patterns and relationships, to make large volumes of information more effectively understood. This core technology drives our product portfolio and mission to transform information into knowledge. We're here to encourage curiosity and deeper understanding of the world.

Based on these core principles, Yewno approaches resource discovery in new ways. First, Yewno works with the full text of any resource. In fact, the more text that is available, the better the concept extraction will be. And second, Yewno is not dependent on human assigned metadata. Through the use of artificial intelligence, it automatically extracts concepts and discerns relationships between them. It is a perfect discovery complement both for full text items we could never afford to catalogue and also as an additional discovery layer on top of full text items we may have catalogued by reaching down into the texts themselves to draw out additional concepts our broad metadata topics would never uncover.

Again, as opposed to linked data, Yewno is a full-text discovery interface; it can analyse any amount of text that you can make available to it whether it be open web resources, full text articles, or material from your digital repository. Yewno semantically analyses that full text, relating similar concepts across very disparate document types. Yewno then presents a graphical interface of those relationships so that you can explore topics in context, allowing you to explore and make new associations between topics your searches reveal. And last, Yewno can give you access to the full text of the data it uses for analysis.

As an example, let's take a look at how Yewno can let us explore the 1957 classic American movie, The Three Faces of Eve.

The Three Faces of Eve



Yewno begins by giving the researcher a brief summary of the topic on the left-hand side of the screen and a graphical summary of all of the topics it has extracted from resources about the topic on the right.

THE JOURNAL OF AMERICAN CULTURE

Explore this journal >

PDF

Info

Original Article

Woman, Divided: Gender, Family, and Multiple Personalities in Media

Katherine J. Lehman

First published: 15 March 2014 Full publication history

DOI: 10.1111/jacc.12107 View/save citation

Cited by (CrossRef): 0 articles Check for updates

Citation tools

View issue TOC
Volume 37, Issue 1
March 2014
Pages 64–73

Woman, Divided: Gender, Family, and Multiple Personalities in Media

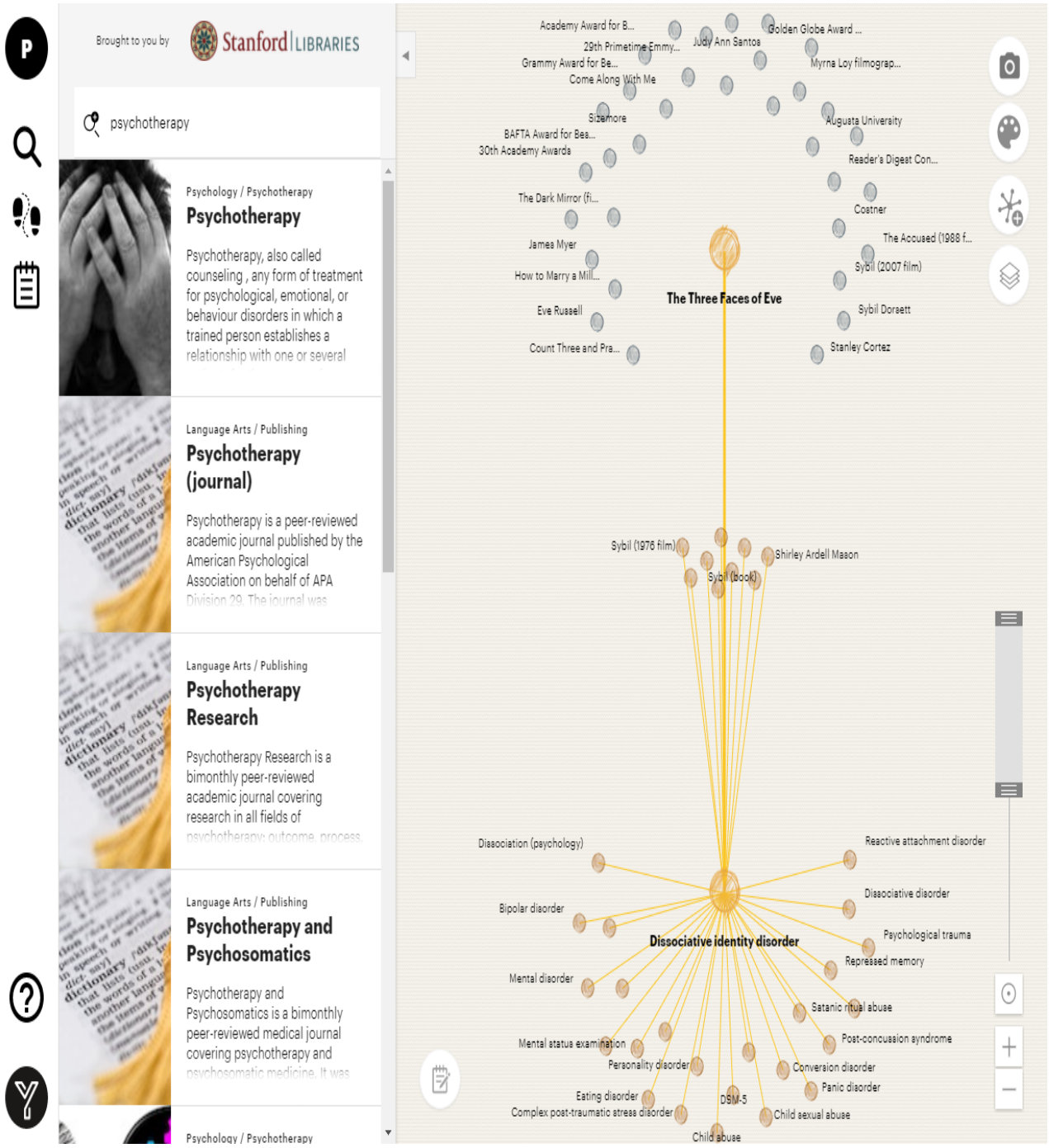
Katherine J. Lehman

In the opening scenes of Showtime's *United States of* Tara (2009–2011), a weary woman faces the camera and drolly narrates her dilemma. Tara (Toni Collette) explains that she is an artist who works on elaborate murals for wealthy clients. She handles her workload and caters to difficult customers, but she can't seem to "micromanage" her teenage daughter's sexual proclivities. At this confession she breaks down, closes her eyes and takes a deep breath. When she reopens them, she has become a different person: no longer the defeated mother, but the brazen teenager "T," who trots off to befriend Tara's daughter. Over the course of the pilot episode, Tara vacillates wildly from one personality to the next, her mental illness playing out not in a therapist's office but in familiar settings like family dinners, ballet but talks with her husband and teenage children about her transitions and treatment.

Although *United States of Tara*'s creators emphasize that the story is a dark comedy, not a documentary, they claim to have consulted DID patients and experts in developing the series ("Diablo Cody"). *Tara* has the potential to foster empathy for mentally ill people and awareness of the long-term effects of childhood abuse, but it may also reinforce popular misconceptions about multiple personalities at time when the diagnosis itself is controversial. Despite its contemporary sensibilities, the series clearly draws from classic media narratives about DID. Tara's initial triad of personalities and domestic setting make the series a contemporary counterpart to the 1957 film *The Three Faces of Eve*, featuring a meek housewife

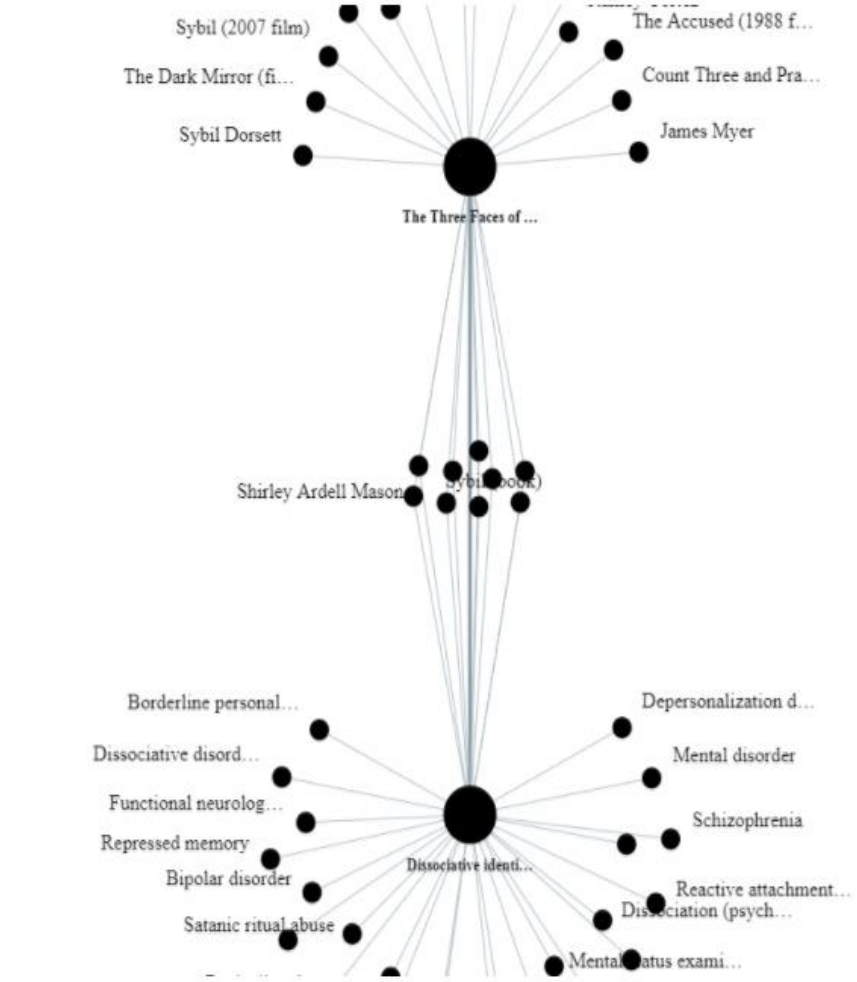
Articles used to support the summary can be opened, explored, and read if desired.

By expanding the bar on the bottom right of the screen, the researcher can expose a deeper level of subject analysis.

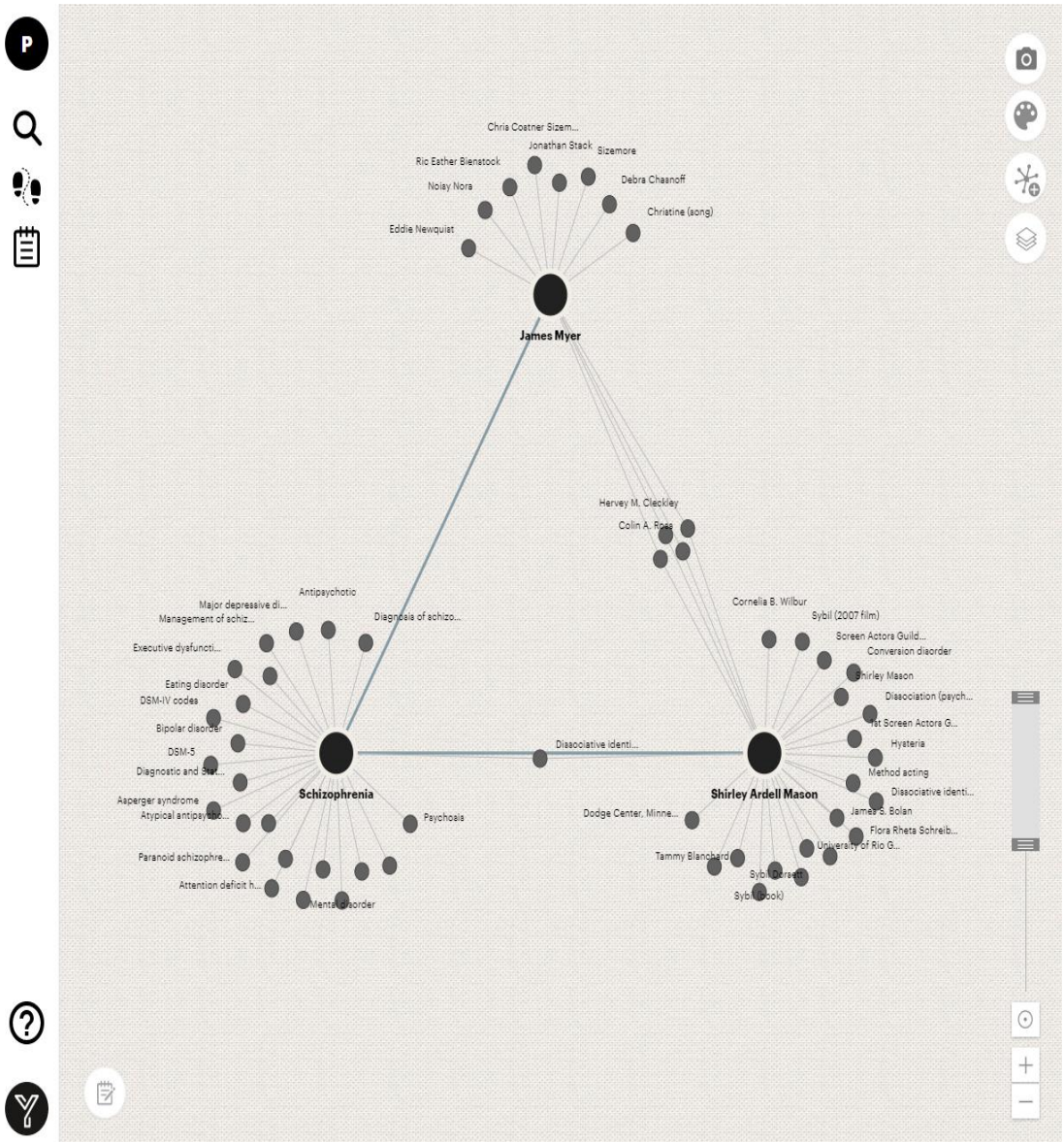


A researcher may also add a second topic of research and see how these two topics relate to each other.

Yewno Research Export



A researcher may also export their report of topics and documents associated with the report.



By adding more search topics, or expanding topics already revealed in earlier searchers, a researcher may reveal a more dense set of relationships.

Yewno, then, provides a powerful complement to more traditional approaches to library resource discovery. In the world of MARC cataloguing, tightly controlled metadata surrogates are created for each resource in a library's collection. These descriptions conform to an exacting set of rules (currently Resource, Description, and Access, or, RDA) and the linking fields, or controlled access points, draw related entities together. The creation of these metadata surrogates is still a one-by-one process for the most part and the creation of the controlled access points limited to a very high level of professional staff.

By shifting to linked data, the language of the semantic web, libraries free their metadata from the confines of a format understood only by libraries (the MARC formats) and allow their metadata to link to the wealth of data on the open Web. The linking that is possible, however, is still limited to the metadata that is pre-assigned by cataloguers. And this sophisticated linking requires all resources to be seen through the same lens, a lens more and more called into

question by an increasingly diverse set of researchers. By not being dependent on human processing, often a constriction in our resource processing cycle, but by making use of artificial intelligence to discern concepts and relationships both within and across limitless unstructured text, Yewno provides a perfect complement to more traditional approaches.

The world of discovery and access is in a time of impatient transformation. Through the conversion of metadata surrogates (cataloguing) to linked data, libraries can represent their traditional holdings on the Web. But in order to provide some form of controlled access to unstructured data, libraries must reach beyond traditional cataloguing techniques to new tools such as artificial intelligence to provide consistent access to a growing world of full-text resources.

References

[1] MARC Homepage. Accessed June 10, 2019. http://www.loc.gov/marc/.
[2] BIBFRAME Homepage. Accessed June 10, 2019. https://www.loc.gov/bibframe/.
[3] Schema.org Homepage. Accessed June 10, 2019. https://schema.org/.
[4] Dublin Core Homepage. Accessed June 10, 2019. http://dublincore.org/.
[5] CIDOC-CRM Homepage. Accessed June 10, 2019. http://www.cidoc-crm.org/.
[6] NACO Homepage. Accessed June 10, 2019. https://www.loc.gov/aba/pcc/naco/
[7] RDA Toolkit Homepage, https://www.rdatoolkit.org/.
[8] PCC Homepage. Accessed June 10, 2019. https://www.loc.gov/aba/pcc/.
[9] CONSER Homepage. Accessed June 10, 2019. https://www.loc.gov/aba/pcc/conser/.
[10] Wikidata Homepage. Accessed June 10, 2019. https://www.wikidata.org/wiki/Wikidata:Main_Page.
[11] Yewno Homepage. Accessed June 10, 2019. https://www.yewno.com/.