THE CHALLENGES OF INCORPORATING GREY LITERATURE INTO A SCHOLARLY PUBLISHING PLATFORM

Alistair Reece

reece@geoscienceworld.org

GeoScienceWorld, USA

Abstract

GeoScienceWorld are in the process of acquiring, converting, and loading a major content repository with a significant amount of grey literature, to be hosted alongside our existing collection of peer-reviewed journals and books in the geosciences. The following issues will be addressed:

What happens when a traditionally scholarly content provider decides to incorporate grey literature into their online content platform?

What are the challenges of preparing the content for publication and discoverability?

How does the presence of grey literature in the database affect cross-search?

How do differing business models find a common home in a unified content platform?

Keywords
GeoSciences, project management, publishing platform, XML, search, business models

---

Introduction
Within the realm of geosciences there is an increasing demand for access to professionally produced, though not peer-reviewed, literature. Such literature comes in the form of reports, both corporate and governmental, meeting abstracts, presentations, and maps, as well as a range of other content types.

The one unifying feature of this content is that while it maintains a high level of integrity within the geoscience community, it has not gone through the academic peer-review process prior to publication.

As one of the leading providers of scholarly content to the geoscience community, GeoScienceWorld saw both an opportunity, and a responsibility, to bring this valuable content to its diverse subscriber base via a single access point, the GeoScienceWorld website.

Background
Established in 2004 to provide a single online source for some of the world's leading scholarly journals and e-books in geosciences. GeoScienceWorld today hosts 47 journals and more than 2 000 books on our platform, from several of the pre-eminent scholarly societies in the geosciences, including the Geological Society of America, Geological Society of London, and the Mineralogical Society of America.

The aim of GSW's founding societies was to bring together peer-reviewed, society led, research on an online platform that would encourage collaboration among the societies in order to benefit the whole collective. This approach allows smaller societies to benefit from being part of a global network of publishers, bring their content to a broader audience, whilst maintaining their independence as societies within the publishing ecosystem.

GeoScienceWorld actively supports the continued research efforts of our member societies, and has channeled more than $35 million back to the societies since our founding 15 years ago.

From the beginning of the platform, our customer base has included corporations and government bodies for whom the body of valuable geoscience content is not limited to peer-reviewed academic journals.

During the migration to our current platform provider, Silverchair, an opportunity arose to acquire a large set of content, of which more than 30% constituted "grey" literature, mainly in the form of meeting abstracts.

With the migration complete, it was decided that incorporating such content into our offering was a valuable, and strategic way forward. Naturally such an acquisition of new content types, there have been a number of challenges raised in the course of planning for the implementation phase of the project, which is scheduled to be complete in early 2020.

These challenges can be summarized as being:

- how to prepare grey literature for loading and publication
- the impact of grey literature on search functionality
- new business models required to support grey literature

Preparing the Content

The process of bringing a new journal or e-book onto the GeoScienceWorld platform is relatively straight forward. Our platform provider, Silverchair, has a stable XML specification for both journal and e-book content, in both instances using a subset of the JATS and BITS tag suites respectively. GeoScienceWorld provides new publishers with the latest version of the specifications and their content vendors create XML files, with associated assets, to be loading through the content loading tool.

The challenge in bringing grey literature into the mix is that there doesn't exist a single authoritative tag suite for handling non-peer-reviewed content. In this circumstance it is necessary to create a custom DTD and the associated XSLT required to get metadata and content into the database. An additional consideration here is that a custom DTD and XSLT is required for each unique content type within the body of content.

Before being able to get to the stage of creating the XML, DTD, and XSLT it is necessary to identify those content pieces which lack the kind of identifiers that are standard in the scholarly publishing world, such as ISSNs, DOIs, and ISBNs. For much of the grey literature in the body of content being brought into the GeoScienceWorld website, such identifiers are either not contained in the content itself or just do not exist.

Impact on Search

With 47 scholarly journals and more than 2 100 e-books on the GeoScienceWorld platform, search is the single most important feature of the website. As such, how the search engine presents non-peer-reviewed content to our users in a manner that reduces potential user confusion while maintaining the current overall user experience is a key consideration in this project.

The GeoScienceWorld platform uses the open source, enterprise scale search engine SOLR to power discoverability throughout the website. Incorporating grey literature into the site requires custom modifications to the SOLR core as well as to the index, adding new fields for the engine to search on. In order to provide the relevant data points to the search engine, the absence of a formal peer-review process has to be indicated through the XML.

Given that the GeoScienceWorld user community consists of both academic and corporate users, it is necessary to clearly indicate the peer-reviewed status of a particular piece of

content. To achieve this aim GSW is implementing two approaches, firstly to introduce a facet into the left rail of the search results page that will allow the user to filter out non-peer-reviewed content, and secondly by using a graphical indicator on the search result that identifies a piece of content as non-peer-reviewed.

The image below shows the search results page as it currently exists. The facet allowing users to filter out non peer-reviewed content will display in the left rail, immediately above the "Format" facet. The image also shows the information presented to the user for each search result. One of the options for the graphical indication that content is not peer-reviewed is to display it on the same line as the Abstract, PDF, Purchase, and Citation Manager options.

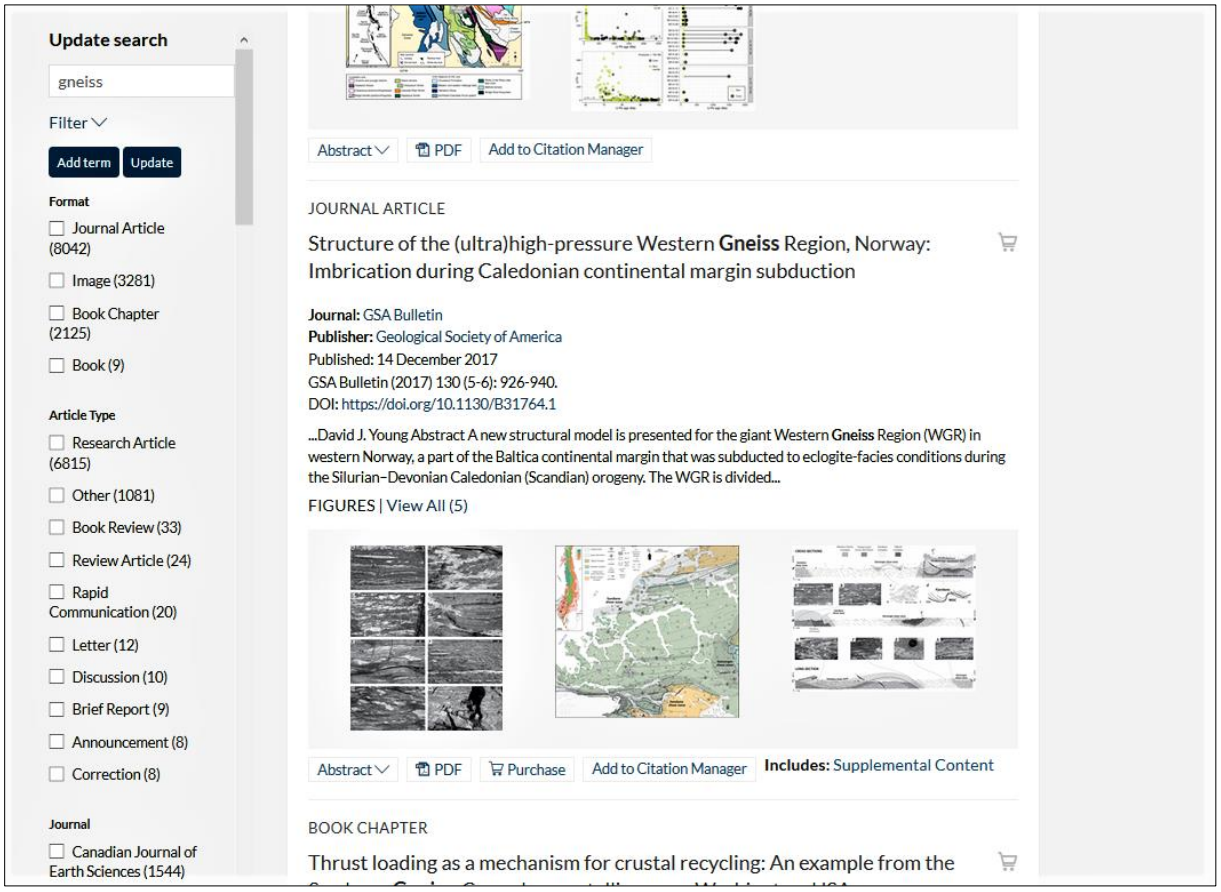

Figure 1: Search results page (GeoScienceWorld)

In reviewing how federated search tools such as EBSCO and ProQuest handle identifying the peer-review status of a piece of content, we noticed that it tended to be hidden as part of a "journal information" drop down. GSW's intention is, however, to make that identification clear to the user without further clicking, thus providing a cleaner user experience. As we move deeper into the migration project, the identification of content as either peer-reviewed or not, and it is attendant impact on user experience will be considered in more detail.

As well as the technical considerations with regards to including non peer-reviewed content into the search experience, it is important that the expectations of the user be accounted for. At present, the majority of users on the GeoScienceWorld platform are students and researchers at academic institutions, whereas the majority of users for the grey literature are coming from the corporate market. The challenge in terms of search here is to make the grey literature accessible to the latter user group while not diminishing the value of peer-reviewed

content in the eyes of the academic market. This consideration is the driving force behind the facet to allow the user to filter out the content that is not relevant to their search.

Our intention is to make the initial search results page contain both peer-reviewed content and grey literature, then allow the user to use the facets in the left rail to further narrow their search as they see fit. We believe this approach has two main benefits, firstly it shows that we trust our users to make their own research decisions, and secondly it potentially brings the grey literature in our corpus to a wider audience. Our aim here is to do nothing that impedes discovery of content, an approach that is widely considered to be best practice for search within a website.

Business Models

In order to support the presence of grey literature on the GeoScienceWorld platform, it is necessary to introduce new business models to the site that support the expected behaviors of targeted customer groups. GSW identified these groups as being corporations, consultancies, government bodies, as well as non-governmental organizations. The current supported business models of subscriptions and pay per view are felt to be restrictive for these target audiences.

Based on GSW's internal research and anecdotal evidence from conversations at conferences and similar events we plan to extend our pay per view functionality to allow bulk purchases of content. Such bulk purchases will be supported through workflow modifications to the classic cart and checkout process through which are currently provided to handle pay per view. In order to cultivate relationships with corporations and consultancies in particular we will also support tokenized purchasing where the customer prepays for a set number of downloads, to be used within a given timeframe.

The current cart and checkout process while technically capable of supporting the purchase of multiple content pieces mitigates against this behavior. When the user places an article or book chapter in the cart, the user is taken to a cart view page that encourages the user to immediately checkout, with no clear method of continuing to browse content. By placing an intermediary step between the functionality to place content in the cart and viewing the cart itself, users will find buying multiple pieces of content less onerous and repetitive. The intermediary step takes the form of a popup that informs the user that content has been placed into the cart and presents them the option to continue browsing or to checkout.

The other form of purchasing and fulfillment that is being investigated to support this content is to allow customers the option to pre-pay for content to be accessed on an ad hoc basis. For example, a customer would opt to buy $1000 of pay per view content and would then have a year to use up that balance. Each time a user associated with that customer is on the site and wants to purchase a piece of content, the price of the content is subtracted from the customer’s remaining balance, with notifications sent to the customer’s administrator informing them of the purchase and new balance.

While this purchase model would not be restricted to only grey literature, it is being considered to support the primary expected users of grey literature, non-academic corporations and consultancies. Such organizations, in general, have moved away from having subscriptions to

journals and ebooks, preferring to cherry pick content and curate their own collections through a knowledge management department. Having such a tokenized offering supports this workflow, as well as streamlining the content expenses process for the customer.

To further support both our academic and corporate customers, especially with the coming of grey literature into the platform, GeoScienceWorld are looking into opening our content to text and data mining tools, either custom built or by implementing an existing tool. This tool would feed into both the extended pay per view and tokenized purchase models.

With these new business models, and purchase methods, we intend to further extend what it means on GeoScienceWorld to purchase content. Traditional access to content, whether through a subscription or pay per view model has given the user the ability to view full text HTML content online or download a PDF version of the same content. Our extended model will allow users the option to download the content's XML files, including metadata, as well as the PDF and any supplementary material.

Conclusion

Any migration of large bodies of content from one platform to another presents a raft of challenges, in the case of peer-reviewed journals and e-books these challenges are largely known and documented as part of a migration process. Migrating grey literature between platforms, especially from a proprietary platform to one of the scholarly platforms such as Silverchair, is very much a case of starting with a few basic assumptions and then discovering the unknowns as the project unfolds.

In pursuing this migration project, GeoScienceWorld have faced challenges related not just to the content itself but how our platform supports, or can be extended to support, the business that surrounds this content. We have been reminded again of the importance of engaging in a thorough discovery process in order to at least have a broad understanding of the major work involved in the project. Such a discovery process though only has lasting value to the project if its findings are accurately represented through requirements documentation, including assumptions, stating work that will not be undertaken, and the acceptance criteria that define the successful fulfillment of the requirement.

While it is important to document the findings of the discovery phase, it is just as important to recognize that requirements can never be fully set in stone, they develop as the project proceeds and more of the unknowns come to light. For this reason, GeoScienceWorld works with our platform partner using an Agile methodology, in this case SCRUM, to constantly be refining the requirements. The ongoing refinement of the requirements allows the software being specifically developed to support the grey literature being incorporated into the site, and for that content to benefit from the features and functionality available, whether that be cross-search, purchase options, or identifying similarly themed content through related content widgets.

As GeoScienceWorld embarks on the next phase of this migration project, actually building out the features needed to support grey literature, we expect most of our assumptions to be challenged, the requirements to need changing many times, and to have a strong partnership

with our platform provider to meet the architectural problems that will likely pop up as we try to make grey literature work in a framework specifically designed for scholarly content.

References

GeoScienceWorld [online]. McLean, VA: GeoScienceWorld, 2019 [Accessed 13 September 2019]. Available from: **https://pubs.geoscienceworld.org/**

Journal Article Tag Suite. *U.S. National Library of Medicine* [online]. Bethesda, MD: National Center for Biotechnology Information, National Institutes of Health, 2019 [Accessed 13 September 2019]. Available from: **https://jats.nlm.nih.gov/**

BITS: Book Interchange Tag Set, 2019. *U.S. National Library of Medicine* [online]. Bethesda, MD: National Center for Biotechnology Information, National Institutes of Health, 2019 [Accessed 13 September 2019]. Available from: **https://jats.nlm.nih.gov/extensions/bits/**