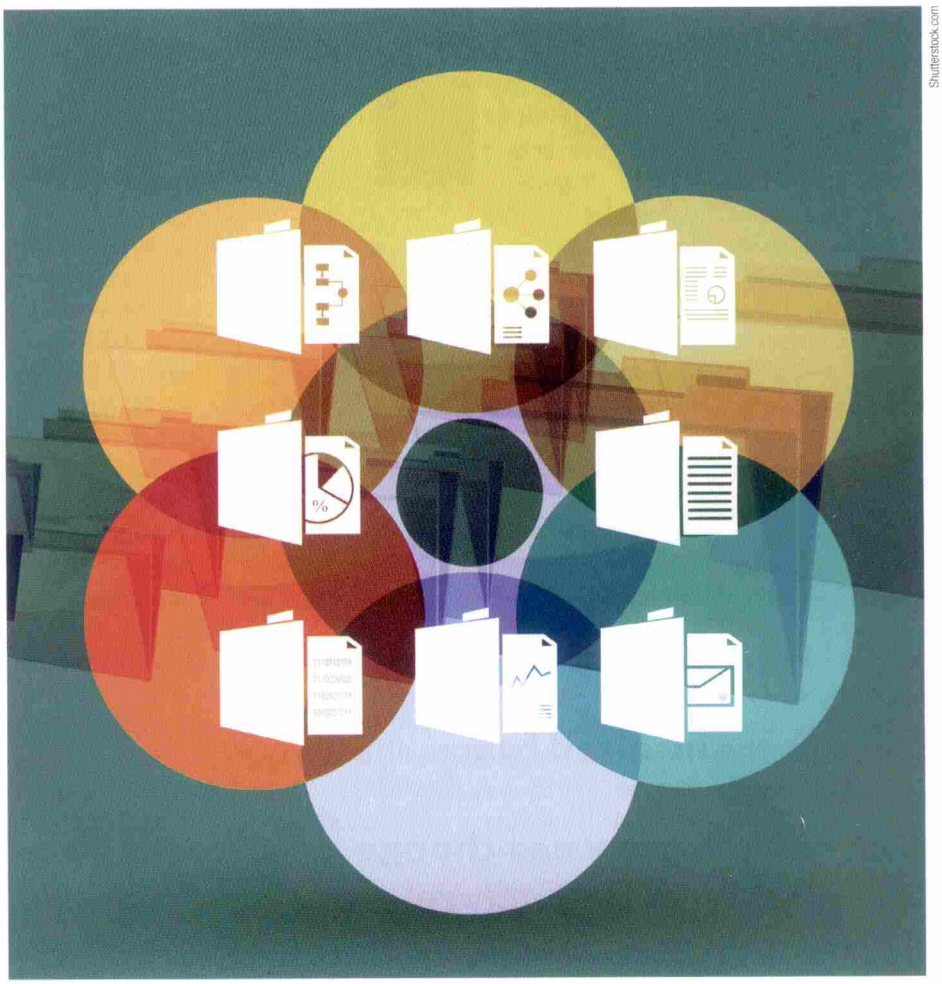
Publishing platforms evolve

Technology plays a major role in how we find and use information today. **Siân Harris** finds out about some of the trends and challenges with publishing platforms



Shutterstock.com

Scholarly content – like scholarship itself – comes in many shapes and sizes. The research process involves information from many sources and of many different types and there is considerable effort going on to bring these different elements together in the most useful way.

When the first big scholarly e-book programmes launched six or seven years ago there was plenty of excitement about putting e-books and e-journals on the same platform so that they could be searched together and this trend has continued.

As Michael Cairns, chief operating officer online at Publishing Technology, observed: 'Over the course of the last few years there has been a recognition that there is benefit to integrate not just books and journals but also conference proceedings.'

Springer has taken this approach with its SpringerLink platform, which the company has recently rebuilt. According to Brian Bishop, the company's vice president of platform development, 'SpringerLink contains journal articles, book chapters and scientific protocols, as well as the "electronic supplementary material" that is published along with these content types, which could be any type of digital media.'

'We store all our full text and metadata in the same database, which is also optimised for search. In the recent rebuild we chose a technology solution that allowed us to consolidate both the database and search features, and eliminated the need for federation or joining of searches across disparate content,' he continued.

He said that consistency between these different types of content is achieved by 'an extremely robust and consistent production process, which results in a very limited set of well-defined outputs (online PDF, XML metadata and full text HTML). This means that the display of content can achieve a very consistent design, giving users a seamless experience regardless of what type of content they are viewing.'

It also, he said, prevents problems in ingesting content because the file formats and metadata received from the company's upstream business partners are already normalised.

Format challenges

However, a range of file formats can pose challenges for platforms, especially where the content being ingested is not just a range of types but also from a range of publishers. As HighWire's director of product management Tara Robenalt observed: 'There are development and support costs for maintaining code branches to normalise various inputs into single forms suitable for hosting with a uniform and scalable code base. Our customers require custom presentations,

and their contractors who supply content to HighWire have varying notions of content standards and quality. In our intake and validation system (HighWire eXpress) public standards are enforced, along with a small set of HighWire-specific source requirements. The challenge is finding and maintaining an appropriate balance between generosity on acceptance, and not letting the normalisation code base drift into an unmanageable or inefficient process.'

Publishing Technology, which also creates and hosts platforms for a range of publishers, has observed similar things. 'On the journals side we always have issues with format for ingest and always anticipate some back and forth to get things how the customer wants,' said Cairns. 'Typically, we specify to publishers that we require XML but there are always exceptions and problems – especially with converting archives.'

And he said that the challenges are greater with e-books. 'Books haven't been online as long, so the issues are more basic. Often we will get a full book PDF that we need to break down into chapters – and metadata is often provided at the book level rather than at the chapter level.'

In addition, he noted that many things such as indexing, endnotes and footnotes work well in print book navigation – but in the online world, especially in content ingestion, these are problematic. 'Even within publishing houses, processes are not consistent,' he observed.

Publishing Technology has two platform products for publishers. IngentaConnect is an aggregator platform. This is often used by smaller publishers and allows content to be searched across many publishers. 'For smaller publishers, being part of that aggregation is good,' said Cairns.

The other approach that the company's customers take is to have standalone, bespoke implementations of Publishing Technology's pub2web platform.

Semantic enrichment

A trend that Cairns is seeing is for publishers to want to do semantic enhancement of their content. In some cases Publishing Technology does some of this, or will ingest semantic enhancement along with content if publishers use a specialist semantic technology company.

Julian Norman, senior product manager at IOP Publishing, expects this trend to continue. 'Semantic technologies offer users an improved reading experience by highlighting the key concepts of content and introducing more advanced linking to additional information. The semantic technologies that are available are also becoming increasingly important for developing new sales propositions and supporting growth into new markets,' he said.

IOP recently launched its e-book programme to sit alongside the publisher's journals on the IOPscience platform. 'Bringing together books and journal content is an exciting opportunity. Not having legacy book content allows flexibility,' he said.

Nonetheless there are challenges in bringing content together on the platform. 'The formats and structures of articles and article files change between journals, suppliers, publishing partner and years, meaning that we have many different configurations of source data to manage,' he said.

'The addition of e-books has given us an additional format of XML to manage but a content format in which we have no legacy content, giving us the opportunities to build new systems around a format we have control over and open up new potential.'

'Publishers need to start thinking of their platforms as more than just a browser-based service and start thinking of them as a data service'

Julian Norman, IOP

He said that book content is stored in the same XML repository as journal content, but that the production system is different. 'IOPscience pulls books out of the repository in a different way to journal content but it is integrated throughout,' he said.

'The more diverse your content gets, the more difficult it gets to manage and bring together for users in a single experience,' he added.

This is a challenge that John Peters, director of GSE Research and Greenleaf Publishing, has also discovered. His company's business includes content from several publishers in the sustainability area and uses the Ingenta Connect platform.

'Being single-point searchable was part of the appeal of going with Ingenta as a partner but we had to fit books to a journals platform,' he said. 'We've had to take in content and run it back through our content system. We break books into chapters with our own metadata, and that means a bit of engineering on our production side.'

Using APIs

One trend that is emerging to help bring together content is the use of application programming interfaces (APIs), which publishers and others are also increasingly opening up to their users and third parties.

As IOP's Norman noted, 'Publishers need to be looking at access to content via more than just the web browser. To be part of any new community projects, APIs need to be available to fetch our content in an easy and open way. If they aren't already, publishers need to start thinking of their platforms as more than just a browser-based service and start thinking of them as a data service.'

He added that, as the publisher uses APIs to build system behind the scenes, he is keen to make these available externally too. 'Return on investment with something like APIs is not immediately apparent – but, if you don't do it, you will be left in the cold,' he observed.

A separate approach

There are sometimes also benefits in keeping different types of publisher content separate, says Richard Padley, managing director of Semantico. He described a project that the technology company is doing for Brill, where the publisher has four platforms that it 'for good reasons is keeping separate.' Reasons for keeping products separate, he said, can include the desire to target specific content and user experiences to specific markets.

Padley explained that Brill wanted each product to stand alone but interoperate seamlessly. Semantico's approach to dealing with this is to build a product called Linking Hub. This, he said, creates context-sensitive links between different sites, achieving the same effect as them all being on one platform but by flexible linking.

He sees this approach as useful in a number of situations. 'For publishers, the approach of putting it all on one big platform only works if the content is quite similar. It's about economies of scale; you don't need to change software much to move from one journal to the next. When you try to include different content types it becomes a bigger technical challenge. You can spend disproportionate time on things that don't fit so well. Publishers

short cutting and trying to make everything look like a journal to use a journal platform can be a case of "square pegs in round holes".'

Many publishers have grown through acquisitions, which brings the potential for greater inconsistencies in content. 'Very few publishers have the luxury to say: "stop everything and we'll do a multimillion-pound platform redevelopment",' said Padley.

In addition, this approach can help with discoverability through search engine optimisation, he added. 'If you link your content together then Google likes it. Publishers are missing a trick if they have lots of silos and don't link them together and show Google the links.'

Going online

For some publishers, the first step is still to put content online. This is a challenge that specialist aggregator Casalini libri takes on with its Torrossa platform. The company brings together humanities and social science full-text content from Italy, France, Spain, Portugal and Greece and often this is the first foray into digital for the nearly 200 small and medium-sized publishers that the company works with.

'The main challenge is to find common parameters and options under which content can be used by the readers. Each publisher would like to adopt personalised options for content delivery. However, the content offered to libraries needs to be streamlined and under homogeneous licences as much as possible,' said Luisa Gaggini, head of e-content and partners relations at Casalini Libri.

Another challenge is to make publisher partners more aware of digital market needs. She said that many small publishers still have a print-focused approach. 'They may not be aware of the importance of changing their mindset and being flexible,' she said. She explained that there can be a fear of losing control of content, concerns about hacking and a feeling that digital is a threat to their small print sales.

'We tell them that avoiding putting content online is not a way to keep sales of print. We need to clarify what the trends in the industry are and not try to stop them.'

However, she added that this is not so simple because, for many small publishers, their small paper sales are their only source of revenue.

'Small publishers may not be aware of the importance of changing their mindset'

Luisa Gaggini, Casalini Libri

Casalini helps by putting partner publishers in contact with an outsourcing service in Italy that digitises materials.

The legacy approach influences the platform's choice of format too. Although Casalini's system can handle EPUB files, content is mainly in the PDF format.

'To produce EPUB versions would require recreation of the article and many publishers are not willing to invest in redoing content when they are unclear of the revenues. PDF is closer to print so it is a sort of baptism into digital,' she explained.

Gaggini said that a key feature of the Torrossa platform is the effort that the company puts into metadata. 'In order to keep clear metadata and to have it homogeneous, we review all the metadata manually. This is a huge task but the only way we can provide a real service.

If you have a publication in three languages you need to make clear that these languages are present in the text,' she explained.

She believes that the platform plays an important role in helping readers uncover specialist non-English-language content. In particular she sees a role in maintaining interest in material written in Italian, a language not widely spoken outside of Italy but that is historically and culturally important (see box: Discoverability).

There are also economies of scale to be realised in joining forces, as several university presses have discovered. Oxford University Press (OUP), for example, has expanded its own monograph platform for use by other presses under the umbrella of University Press Scholarship Online (UPSO).

As Kurt Hettler, director of GAB sales and marketing at OUP described, 'the idea behind UPSO is to disseminate the best monographic scholarship to the widest possible audience. There aren't any additional challenges as a result of being a multi-publisher platform that there wouldn't be if were just Oxford content; our platform was built to be able to handle changes across the broad range of products we produce.'

He explained that the UPSO approach is to only accept printer-ready PDF files or InDesign and other source files. 'We then convert this to our proprietary XML format. For the ingestion of metadata, we have created a comprehensive template that partner presses populate. That, among other things, helps us identify more complex titles such as those with lots of figures, tables, illustrations, and non-Roman characters.'

Bringing together content from a range of presses brings cross-searching benefits, he said. 'The real strength of UPSO is its deep searching capabilities. The XML format we originally created for Oxford Scholarship Online enables us to tag data at a very deep level, ensuring that search results are far more robust than those that would be returned on a simple text search of a PDF file.'

He added that the team is in the process of linking all citations and bibliographic entries to other titles on the platform too. In addition, the press has recently added the Oxford Index link as an underbar to every page of every Oxford online product. 'This enables us to bring in users into UPSO from all of our online products, including the millions of users of our various journals,' said Hettler.

Discoverability

One of the key requirements of a publishing platform, whether built by a publisher or a partner, is to help discovery of content. In this, working together can help.

Luisa Gaggini of Casalini Libri noted, 'We try to use as many discovery channels as possible. We can be targeted with federated search Z39.50 protocol, and have an OAI-PMH server for metadata harvesting. We also distribute metadata to all discovery services and supply customers with MARC records to all titles supplied.'

John Peters of GSE and Greenleaf agreed on the importance of this, adding that his company's content, on Ingenta Connect, is exposed to the likes of Summon, Primo and EBSCO Discovery Service. 'Small publishers have to be very aware that you need to work with the big players. The days of publishers saying they will be in full control of their content are past. I would prefer to have ProQuest as a friend than someone who doesn't know or care about us.'

And this is what users want too. Peters recalled that, when GSE launched, 'libraries said to us: "Give us as much flexibility as possible and make sure it works with discovery tools".'

The future

As challenges around ingestion, formats and metadata are addressed, new challenges emerge. Looking to the future, there is the possibility of more different types of content.

What stage this is at varies by discipline and publisher. 'Enhancements, like video and interactive features, will probably be what shapes the future of e-books, and having UPSO in XML format really makes us future-ready, in a much more robust way than PDF replicas of books,' said Hettler.

Gaggini at Casalini said that she and colleagues are keeping an eye on this trend but do not see a major drive for this at the moment. 'We don't see much video coming out with papers and we don't think it will happen soon. The cost of producing such interactivity is huge against low returns, at least for non-English language HSS research publications.'

Cairns also noted that the majority of content that Publishing Technology handles for publishers is traditional, text-based material. 'Video and audio are something of an exception,' he said. 'We could probably be very sophisticated but we are not really being asked to do this at the moment.'

This trend does pose additional challenges, however, according to Robenalt of HighWire. 'Files that accompany the content source – for example a flash file, video file, or other kind of data supplement – require that the hosting platform can ingest the files, associate them to the appropriate piece of content, and deliver them on the site in the location preferred by the publisher. HighWire allows the publisher to do all of these things, most recently allowing the publisher to define in the XML source where to link to a file from within an article, with some configuration options to define the interaction,' she said.

But industry standards are still lacking, according to Springer's Bishop. 'The largest problem we face is the implementation of standards in how we decide to support new features.' He pointed to the example of the lack of broad support amongst browsers for including Scalable Vector Graphics (SVG) or MathML. 'When you produce over 350,000 content items per year it is extremely important to ensure that any alterations or additions to that process are well documented and supported,' he argued.

Thinking about such issues is important as interest in including different types of content grows. 'Authors increasingly ask questions of publishers' ability to publish rich, interactive media in an accessible way. They see that they can embed a wide range of content within their web pages and are pushing publishers to offer the ability to embed this content in their articles rather than include it as a hidden "supplementary" file,' observed IOP's Norman.

With any looking forward there is also a need to look back. 'The biggest challenge with platform functionality is understanding the complexities of introducing the functionality retrospectively. At any one time we want to produce our content in the best format available so we look at introducing functionality for that format and then backdating it if possible. We have invested in XML transforms for converting older XML to new DTDs on-the-fly in order to support new functionality for older content, said Norman.

Such efforts should help researchers and others data mine content. 'We need to make sure stuff becomes part of the online ecosystem. We have a growing responsibility to link up this information for the benefit of researchers,' he added.

The utopian vision is summed up by HighWire's Robenalt: 'Publishing platforms will continue to find ways to provide a seamless experience for readers, making it easier to find all of the content available from one place, share content with peers, and integrate with other tools that readers use to discover, consume and manage the content they are interested in.' ■