Interview with Henning Schoenenberger

*Director, Product Data and Metadata, Springer*

Henning is responsible for metadata globally for Springer, one of the world’s largest and oldest STM publishers. Springer has long been an innovator in the digital arena, pioneering the use of SGML and XML, publishing journals online, embracing open access, and—most notably in the context of this initiative—digitizing their enormous backlist of some 100,000+ books in a broad range of disciplines in their SpringerLink platform.

Because his work focuses almost entirely on metadata for a publishing program that encompasses many thousands of books and journals with decades of legacy content, Henning has an uncommon perspective and depth of knowledge in this area.

He has a very comprehensive and forward-looking view of metadata. He pushed back on the focus on “pain points” by saying that he doesn’t consider metadata to be a problem statement; rather, “it’s a central aspect of what you do as a publisher.” One of my favorite quotes from the interview:

*“Metadata is not to be solved, it’s to be saved.”*

If a publisher invests in metadata activities (e.g., CrossRef), it is likely to pay off. His philosophy is to “be as transparent with metadata as possible, flood the market with it, across the world. Same with Linked Open Data and the Semantic Web.” This is not just idealism, it is practical: “This will all benefit Springer in the long run.”

Metadata should be considered an ongoing issue, and should be understood to be always evolving. The development of the DOI and CrossRef was only one part of the picture, one aspect of the ecosystem. That is working well for what it was originally designed to do, but now it has evolved to include FundRef [a standard for metadata that identifies the funder of the research on which a publication was based, a new initiative from CrossRef], which is an issue that publishers must now deal with, incorporate into their systems and workflows and publications.

Especially from an STM publisher’s standpoint, you must have very good metadata to position yourself properly in the marketplace and in the scientific/scholarly dialogue.

An example of how significantly the landscape changes: right now, research data is considered “an appendage” to a publication. But who knows to what extent it may become core of STM publishing? Currently, “research data is a mess: there are hundreds of different formats out there.” You can’t always know where things will go, and when. The lesson is that you need to capture and maintain all possible metadata, even if it may not have a demonstrable payback right now: you will almost surely need it in the long run.

As an example of how this philosophy and strategy paid off, Springer needed to act fast to implement ONIX-PC, the subscription format of ONIX [a complement to ONIX for Books that is currently mainly oriented to journals]. Because they had such good metadata, they were able to implement it on a very large scale in only three months: “It turned out to be quite simple, we needed some IT and some project management and a converter and we were done.” Other publishers struggle with this because they lack the necessary metadata.

On the subject of ONIX-PC, he said that at the ICEDIS meeting at Frankfurt [ICEDIS is EDItEUR's special interest group for subscriptions, serials and library supply], he advocated renaming the “PC” suffix from the current meaning of “Price Catalog” to “Product Catalog.” He wants it to cover Open Access, hybrid products and packages, which it currently can’t accommodate. He also wants to be able to use it to communicate with the ISSN catalog, etc.

This brought up another important general issue:

*“Metadata in the past was tailored to two different markets, trade and library. THIS CANNOT CONTINUE.”* [Yes, I heard the caps over the phone. ;-) ]

Concerning metadata these worlds are coming together rapidly. The separation in their metadata models and management systems—how they express subject classifications, the identifiers they use, how they handle linked open data, the fact that they are typically managed by separate teams, and have separate business models—all this needs to converge.

*“Functionalities and formats (e.g., schema.org) need to cover library and trade under the same umbrella.”*

Another huge issue he emphasized was interoperability.

As an STM publisher, Springer has MARC records; in fact, they have both OCLC MARC records and their own MARC records. The Springer-flavored MARC records have URIs that lead back to the Springer repository, as well as three or four classification schemes, a full TOC, Dewey and LOC classification, and proprietary metadata.

The inconsistencies in MARC records is a problem; there must be reliable readability by different facilities. BIBFRAME is going in the right direction. [BIBFRAME is a proposed successor to MARC, mainly developed and promoted by the LOC and Virginia Tech.] But academic libraries often have limited budgets, and moving away from MARC is not trivial. “This is 80% of the story: it works okay; workarounds are common.” BIBFRAME needs to make the case for libraries for why it’s justified.

He mentioned that he was approached to collaborate with schema.org to model serial and journal metadata. He recommends not reinventing the wheel or developing Yet Another Standard: “How ONIX-PC models serials is excellent. If schema.org is looking for a deep object model for serials data, they should look at ONIX-PC.”

He pointed out that EDItEUR is based on a trade supply chain model and perspective, dominated by wholesalers. But these players are “only part of the story.” Publishers are increasing end-user sales that don’t involve the intermediaries for which the EDItEUR standards have been designed. ONIX is built for the current supply chain model; EDItEUR needs to move beyond this current focus.

This is especially relevant for STM because STM publishers have the brand recognition and the capabilities to be able to sell direct. You can go to SpringerLink today and buy an article or even a book chapter.

Finally, he placed a strong emphasis on rights: “Some still say ‘Metadata is a challenge’. But that is news from yesterday. The current challenge is Rights metadata.” A few technology companies provide decent rights modules. But there is no thoroughgoing interoperable rights standard. “This gives us a great opportunity to tie up loose ends and get it right.”

He would very much welcome the W3C setting up a WG to work on rights, and making rights INTEROPERABLE. Publishers would appreciate this. He mentioned the enormous task that was involved in expanding SpringerLink to accommodate all of their books, which have been published over many decades.

*“Clear rights metadata is especially important now that everything is visible: once you publish something digitally, retractions are a complex matter. This can be life-or-death for a small publisher which hasn’t done its homework.”*

In an abundance of caution, although they did digitize the entire backlist going back to the late 18th century and comprising more than 100.000 books, Springer is highly scrupulous not making public any content for which it has not cleared the rights.

This was an enormous project: it involved hundreds of thousands of contracts from which rights metadata had to be extracted. And these exist in various forms in various jurisdictions.

In conclusion, what Henning characterized as “the Big Data promise”:

* Linked Open Data
* Named Entity Recognition
* Robust Identifiers

With these, it would be possible to realize Henning’s vision:

*“In a few years, robots will do research for you. Good metadata will make this possible.”*