第 2 回 SPARC Japan セミナー 2008

「学術出版と XML 対応 - 日本の課題」

事例紹介 1

物理系ジャーナルの場合

太宰　達三（物理系学術誌刊行センター）

国立情報学研究所

講演要旨

物理系ジャーナルの場合

科学ジャーナルの分野では，XML による組版が海外では主流となっている。それは，いわゆる one source multiuse の概念に沿ったものであり，彼らは XML データを活用するための戦略を持っている。日本でも，XML 組版の重要性を訴える声が上がってきているが，一歩間違えると「XML データを作ることが目的」化されてしまう恐れがある。この講演では，物理系ジャーナルの製作形態の変遷を紹介するとともに，(1) 我々は一体どのような戦略を持ってデータを作成するのか，(2) そしてそれは XML でなければ実現不可能なのか，(3) 日本の現状に即した現実的な解決方法は何なのか等を，物理系ジャーナルの製作担当者の立場から述べる。

講演者プロフィール

太宰　達三

物理系学術誌刊行センター

日本物理学会／応用物理学会　物理系学術誌刊行センター業務部長。主に日本物理学会発行の Journal of the Physical Society of Japan と応用物理学会発行の Applied Physics Express, Japanese Journal of Applied Physics の刊行事業を全般的に扱う。投稿審査システムやオンラインジャーナルシステム，構造化文書の仕様策定などにも関わってきた。1992 年，前身の応用物理学欧文誌刊行会に加わり，2008 年から現職。

基本スタンス

我々が出している物理系ジャーナルの場合を事例として紹介します。

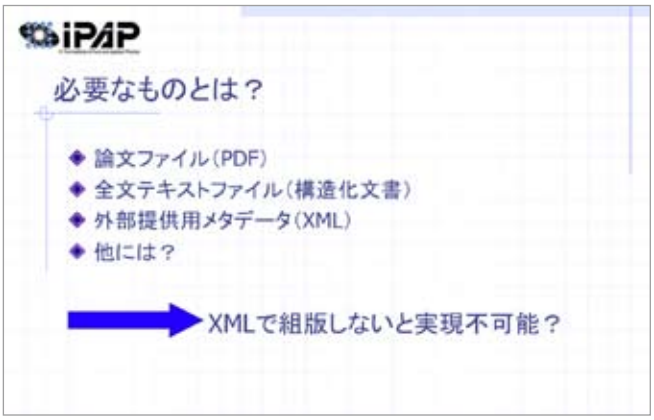
歴史的な流れとして、活版、電算写植、DTP、LaTeX、DB 組版、その他いろいろありますが、こちらが必要なものが問題なく適正な価格で作成できれば、何で作っても構わないというのが、私の基本的なスタンスです。

その場合に必要なものは、論文ファイル（PDF）、全文テキストファイル（構造化文書）、外部提供用メタデータなどです。

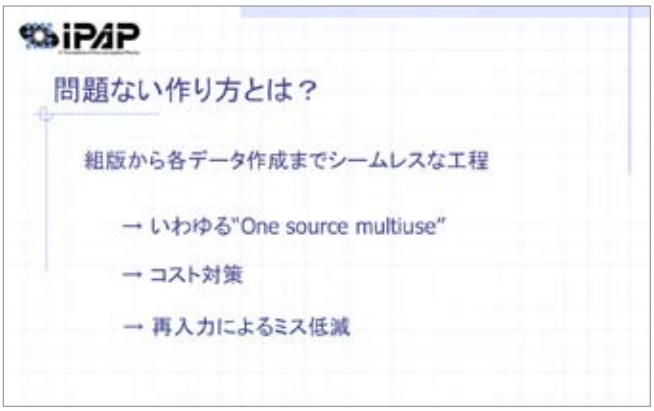
では、問題のない作り方とは何かというと、基本は組版のデータから各データまでシームレスな工程を持てるものということです。

このためには、"One source multiuse" が基本です。

（図 1）



（図 2）



（図 3）

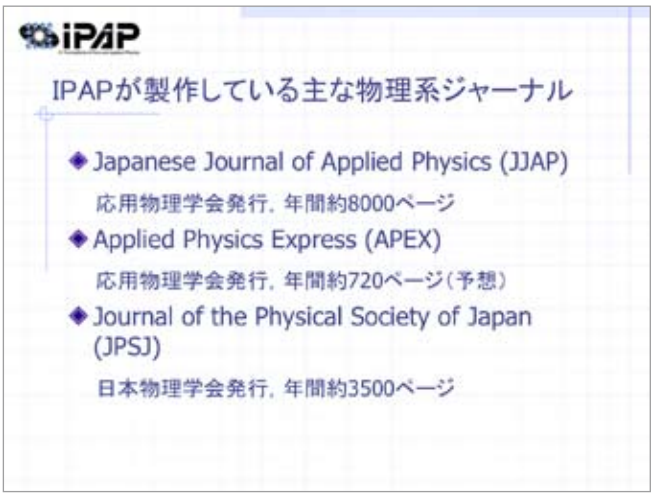
IPAP が製作している主な物理系ジャーナル

IPAP が製作している主な物理系ジャーナルを紹介します。（図 4）

○『Japanese Journal of Applied Physics（JJAP）』応用物理学会発行 / 年間約 8000 ページ

○『Applied Physics Express（APEX）』応用物理学会発行 / 約 720 ページ（2008 年創刊）

○『Journal of the Physical Society of Japan（JPSJ）』日本物理学会発行 / 年間約 3500 ページ



（図 4）

JJAP/JPSJ

ここでは JJAP と JPSJ を例に説明します。

JJAP と JPSJ は、電算写植の後に LaTeX を導入しました。（図 5）

まだ現在のようにネットが普及しておらず、本文のデータを活用するといった時代ではなかったので、LaTeX を導入したのは、単純に著者のデータを版下として使えるようにしたいという、コスト対策です。

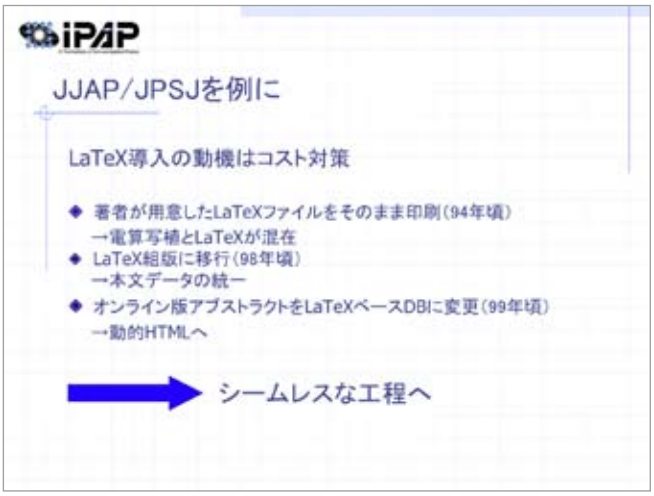
初めは、著者が用意した LaTeX ファイルをそのまま印刷しており、電算写植のデータと LaTeX のデータが混在していましたが、LaTeX 組版に全面的に移行したのが 1998 年頃です。これに 4 年かかりました。

ここで本文データが統一され、それにともない、オンライン版の HTML のアブストラクト部分を LaTeX ベースのデータベースに変更しました。

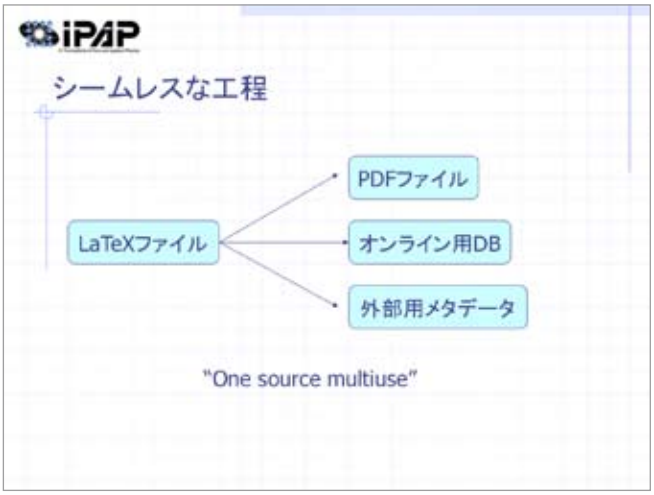
それまでは、HTML ファイルを変換して作って、Web にアップしていたので、どこか少し間違えるともう一度初めから作り直さなければなりませんでした。例えば、100 論文あったとして、99 論文目に何かエラーが起きると、また 1 番目からやり直すという、運用が非常に困難なものだったため、オンライン版アブストラクトを LaTeX ベースのデータベースに変更し、アクセス時に CGI を使って HTML を表示させ、いわゆる動的な HTML にしたのです。

（図 6）は、そのイメージです。これで見かけ上はシームレスな工程になりました。

LaTeX ファイルから版下、つまり論文の PDF ファイル、オンライン用の DB ファイル、外部提供用メタデータファイルになります。これで左側の One source が右の 3 つの multi source に、“One source multiuse” になったことになります。



（図 5）



（図 6）

LaTeX の問題点

しかし、LaTeX にもいろいろな問題があります。（図 7）

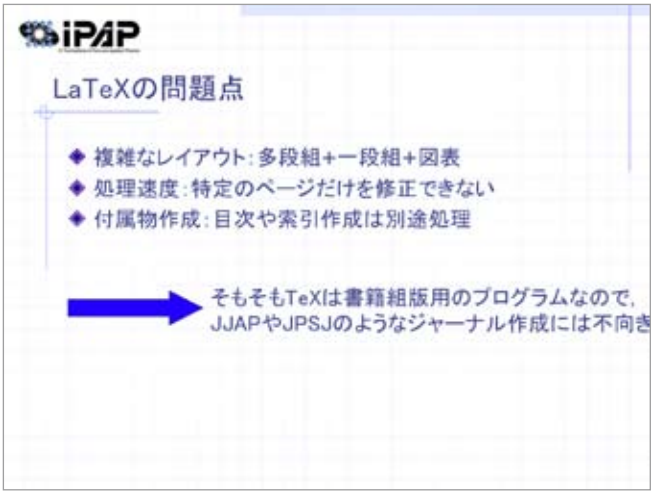
データの問題というよりは、いわゆる組版上、デザイン上の問題です。

一つには、基礎物理で扱う数式の多さです。

JPSJ と呼ばれている日本物理学会が発行しているものは、数式が多く、基本的に 2 段組みですが、途中で 1 段組みになったり、なおかつ、図表がそのページにたくさん入ったりしますので、LaTeX ではレイアウトが非常に難しいのです。

また、一つには、処理速度の問題があります。

一時期の DB 組版ソフトも同じかもしれませんが、特定のページだけを修正できないのです。例えば、1 論文が 10 ページあったとして、9 ページ目のある一部分だけを直したくても、結局、もう一度作り直し（再コンパイル）になってしまうわけです。このため、修正が入ると非常に手間も時間もかかります。



（図 7）

さらに、付属物作成の問題もあります。ここで付属物と呼んでいるのは、いわゆるその号の目次や索引などです。

導入から少し時間がたって、スクリプトを使ったりしていろいろできるようになりましたが、導入当時は、目次や索引は別途処理していたため、データの同期がなかなか取れませんでした。

例えば、論文のタイトルを直したら、目次も自動的に直ると良いのですが、なかなかそういうわけにはいきませんでした。

LaTeX と TeX を区別していますが、そもそも TeX 自体が書籍組版用のプログラムなので、JJAP や JPSJ といった論文、特に、面倒なレイアウトを含んだジャーナルを作るには不向きと考えるようになりました。

新組版システムに移行（2001 年）

そこで 2001 年に新組版システムに移行しました。（図 8）

これはマークアップ組版、要はタグです。

XML でも何でもいいのですが、タグ言語の組版と WYSWYG が組み合わさっているものです。WYSWYG とは「What You See is What You Get」、つまり、見たままの通りに印刷されるというもので、本当の論文のイメージを直すと元のデータも直るという、TeX では考えられなかった仕組みです。

LaTeX 対応とは、数式が plainTeX をデフォルトしているので、TeX のファイルを取り込んで組版に持っていくシステムです。当時は、SGML 対応をうたっていましたが、今は XML 対応になっています。目次や索引といった付属部も、データの同期を取って作れます。

これによって、処理速度は大幅に向上しました。

結局、新しい組版システムに移行したのは、データベースやメタデータのファイルの問題ではなく、単純に、ジャーナルを運用するに当たっての時間とコストの節約が主な理由です。実は、今でもこれを使用しています。

基本ファイルは LaTeX

基本のデータは、今でも LaTeX を使用しています。（図 9）

ファイルを XML にしていない理由は、一つには数式があります。

「MathML の敷居の高さ」と書いていますが、MathML に限らず、人の手で書けるものではないので、修正が入ったときに、MathML を直接いじって修正するのはなかなか難しいのです。XML に比べ、LaTeX は同じタグの言語であっても、何とか人の手が入れられる、人の手で書けるという面があります。

専門家にとっては必ずしもそうでないかもしれませんが、我々の業種レベルでは、LaTeX が持っている構造くらいがちょうど良いのです。

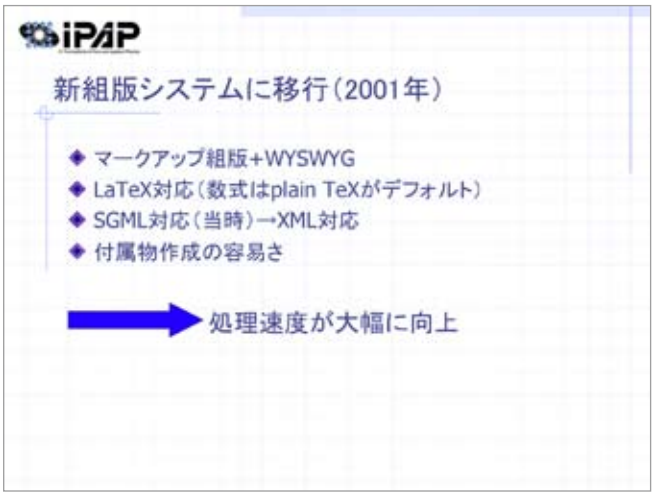
また、ある種の変換ソフトで LaTeX から XML ファイルを作ることができるので、初めから敷居の高い XML で作るよりは、LaTeX で持っておくほうが良いということもあります。

日本のジャーナルの中では大きいほうだと言われている我々でも、年間 1 万 5000 ページ程度で、海外の大手の出版社や大きな学会が行っているような、インドでデータ作成とか、いわゆる学術的プランテーションによる大量生産は、なかなか難しいのが現状です。ですから、人海戦術で取りあえず何とか作れるものとしては、自然な言語に近い LaTeX が一番良いと考えています。

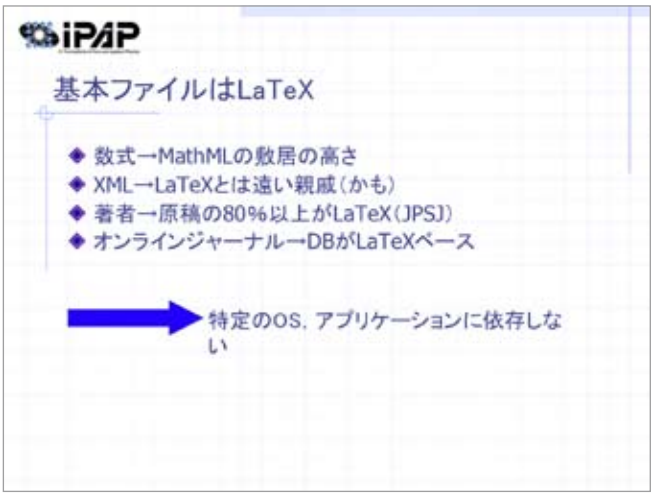
Word から XML ということも、もちろん分野によっては非常に有効だと思いますが、物理、特に基礎物理、数学などにおいては、Word の数式エディタで、1 論文につき 100 以上の大きな数式を研究者が書けるとは思えません。そういう意味でも、今のところは、基本的なデータは LaTeX で持つのが良いと考えています。

JPSJ の場合、著者の原稿の 80%以上が LaTeX です。しかし、それが必ずしもスムーズな工程につながっているかというと、そうともいえません。

新しい組版システムは、正規化、つまりある決まりにのっとっ



（図 8）



（図 9）

た TeX しか読み込めません。しかし、LaTeX は、自分でいくらでもタグを作れるため、必要のないタグがいくつも作られていることがあります。例えば、ある著者の LaTeX ファイルをエディタで開くと、マクロと呼ばれるタグが 100 行ほどもあり、そのタグを使用しているのかを調べてみると、使われていないことが分かり、そうした必要のないタグを削除するなど、データをきれいにする工程に非常に時間がかかるのです。原稿をまったくいじる必要がない場合は良いのですが、原稿によっては逆に手間がかかることもあります。

また、基本データを LaTeX にしている最大の理由が、私ども独自のオンラインジャーナルのデータベースが LaTeX ベースだということです。これが XML ベースに変わらない限り、全文を XML で持つメリットがあまりないのです。

ただし、これは現状の話であって将来的にどうするかはわかりません。

それと、特定の OS やアプリケーションに依存しないということも理由の一つです。

例えば、Word は、デファクトスタンダードだと考えられる分野ではあると思いますが、バージョンアップなどで常に最新のものに追いかけていかないとなりません。ただ、出来上がった XML については、LaTeX と同様、特定の OS やアプリケーションに依存しないので、その辺をどう考えるかだと思います。

XML 組版への移行

将来的に XML 組版に移行するかどうかは、全文データをどう活用するかによると思います。（図 10）

XML を使ったメタデータが必要なのは、いわゆる 2 次データ業者向けであるとか、あるいは CrossRef などの業界団体に渡すメタデータのためだけです。そのメタデータ、いわゆる書誌情報ですが、そのデータを作るためだけに、全文データを XML で作らなければならないかというと、そうではないと思います。

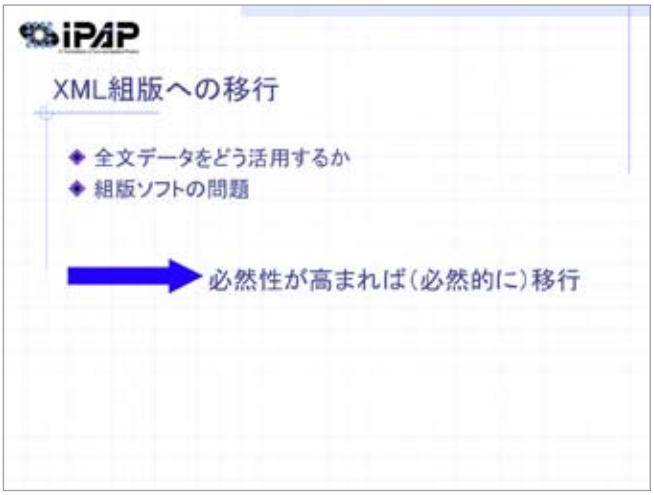
ただ、皆さんご存知のように全世界的に論文の盗作がかなり問題になってきており、それをどうやって未然に防ぐか、見つけるかといったときに、全文データが必ず必要なわけですが、その全文データを独自の形式で持っていても仕方がないわけです。

盗作は、例えば JJAP の論文を盗作した場合に JJAP に投稿する人はあまりおりません。他のジャーナルに出ているものを盗作して、自分のところに出すわけです。もっと言えば、物理の分野では、Physical Review などを盗用して、我々のところに投稿してくる方が少なくありません。こうした盗作の対策には、出版社や学会が共同で類似検索などの仕組みを作る必要があります。そうした流れ、あるいは動きがかなり強くなってきた場合に、それなら初めから XML にしたほうが合理的だろうという話が出てくるかもしれないとは考えています。

それから、組版ソフトの問題があります。

私どもでは、年間約 1 万 5000 ページ、なおかつ英語で、面倒な数式もあるジャーナルを製作しています。分野によってはかなりプレーンなテキストの論文も多いと思いますが、そのXML ベースの英語の論文を得意とするようなソフトを導入して、いつペイできるかというコストの問題です。

例えばアメリカの化学会や、物理学会のように、大きな団体あるいは学会で何十誌もジャーナルを持っているところは良いでしょうが、日本の単体の学会のジャーナル製作用にわざわ



（図 10）

ざ導入するだろうかという疑問が残ります。

しかしながら、全文を XML で組版する必然性が高まれば、移行していくでしょうが、まだそこまで必然性が高くないときに、無理やりやるといった先駆者的な役割を担うことは考えていません。もちろん誰かがやらなければいけないことだとは分かっていますが、今回はフォロワーになろうと思っています。

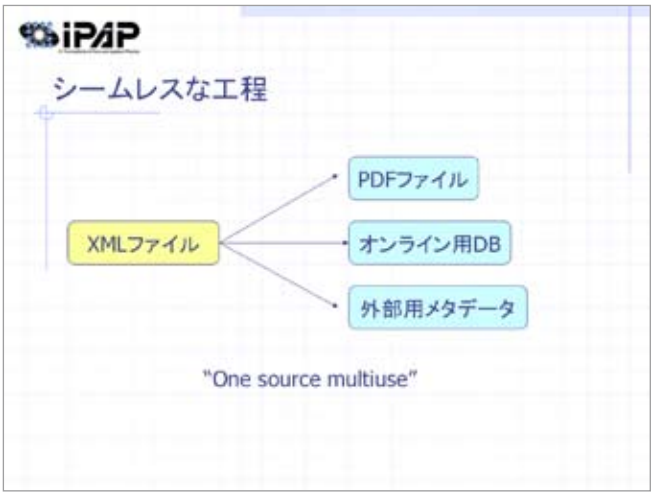
では、XML へ移行した場合、シームレスな工程がどうなるかといいますと、今は、LaTeX を基礎にしていろいろ活用していますが、この LaTeX のところが XML になるだけのことだと考えています。（図 11）

LaTeX をベースとした、いわゆる “One source multiuse” が XML ベースに変わるだけだということです。

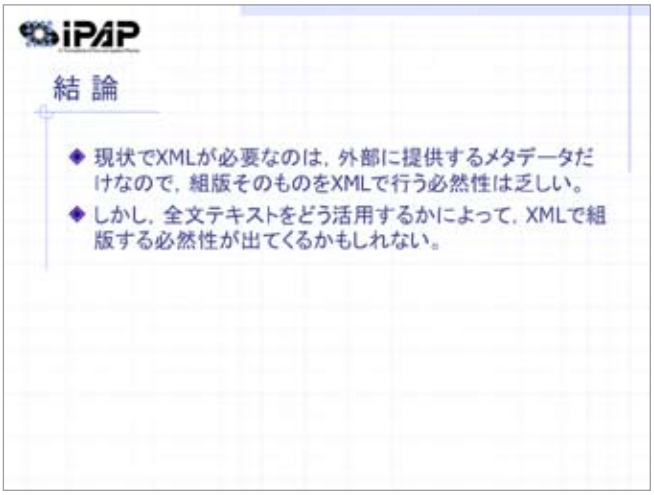
結論として、我々のようなジャーナルの場合、XML が必要なのは外部に提供するメタデータだけなので、組版そのものを XML で行う必然性は乏しいと言わざるを得ません。（図 12）

いろいろ困難がある現状でそれを無理やりやろうとしても、なかなかうまくいかないと思います。ただ、前にも少し触れましたが、全文テキストをどのように活用するかというときに、XML 組版をしたほうが良いという必然性が出てくるかもしれません。ですから、XML を作るための XML はあまり意味がなく、必要なものは何かというものの積み上げによって、XML でないと駄目となったら、多分そうなるということです。

欧米では、ほとんど XML になっているので、やはり XML でやらなければならないのではないかと考える方もいるでしょうが、前述のように、海外の大手出版社のやり方を、そのまま日本の小さい学協会ジャーナルに当てはめるのは、相当無理があります。その無理がなくなるまでは、このスタンスで進んでいこうと思っています。



（図 11）



（図 12）