SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

第2回 SPARC Japan セミナー2008
「学術出版とXML対応-日本の課題」

2008年6月24日（火）

13:30-17:00

国立情報学研究所(NII)

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

経緯と若干のおさらい

Publisher side

- 2007年第3回SPARCセミナー
- 欧米に比較してSGML-XML対応が遅れた日本
- NLM-DTD　スタンダードDTDの登場
- 日本で本格的なXML出版は可能か（TeXの利便性）

DB-solution side

- データベース構築の中でのSGML、XML
- マニュアル、辞書等での実績
- 学術出版への応用例も
- 細かい組版まで対応可能か

両者は近いようで遠い存在

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

今日の内容と理解を助けるキーワード

- 1部：XMLの専門家から学術出版を語ってもらう
  - データXMLの実績から平文ドキュメントXMLへ
  - 後半（審査後）の構造化から前半（投稿段階）での構造化
- 2部：現状のメタデータ出版の事例紹介
  - 数学・物理系　TeXからXML　非常に効率的な出版体制
  - 化学・生物系　WordからXML　日本ではまだチャレンジ
- 3部：ディスカッション
  - 1部と2部の理解を深める
  - 両者のギャップ、距離感の原因はどこにあるか
  - 機関レポジトリの活動からはどう見えるか
  - メタデータ出版の将来像

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

ご注意（参加者が多岐に渡っているので）

- ここでのXMLおよびメタデータ出版は学術ジャーナルのものを指す
  - 辞書などのデータコンテンツ主体ではない
  - 教科書、小説などの書籍ベースでもない
- One style does not fit all
  - 参加者の事情にあったメタデータ（作成）の理解を深める
  - XMLを中心に据えてはいるが、XMLを作るためにXMLを作るのではない

第2回 SPARC Japan セミナー2008
「学術出版とXML対応-日本の課題」

化学系ジャーナルの場合
-WordからXMLを作る試み-

2008年6月24日（火）

日本化学会　林　和弘
hayashi@chemistry.or.jp

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

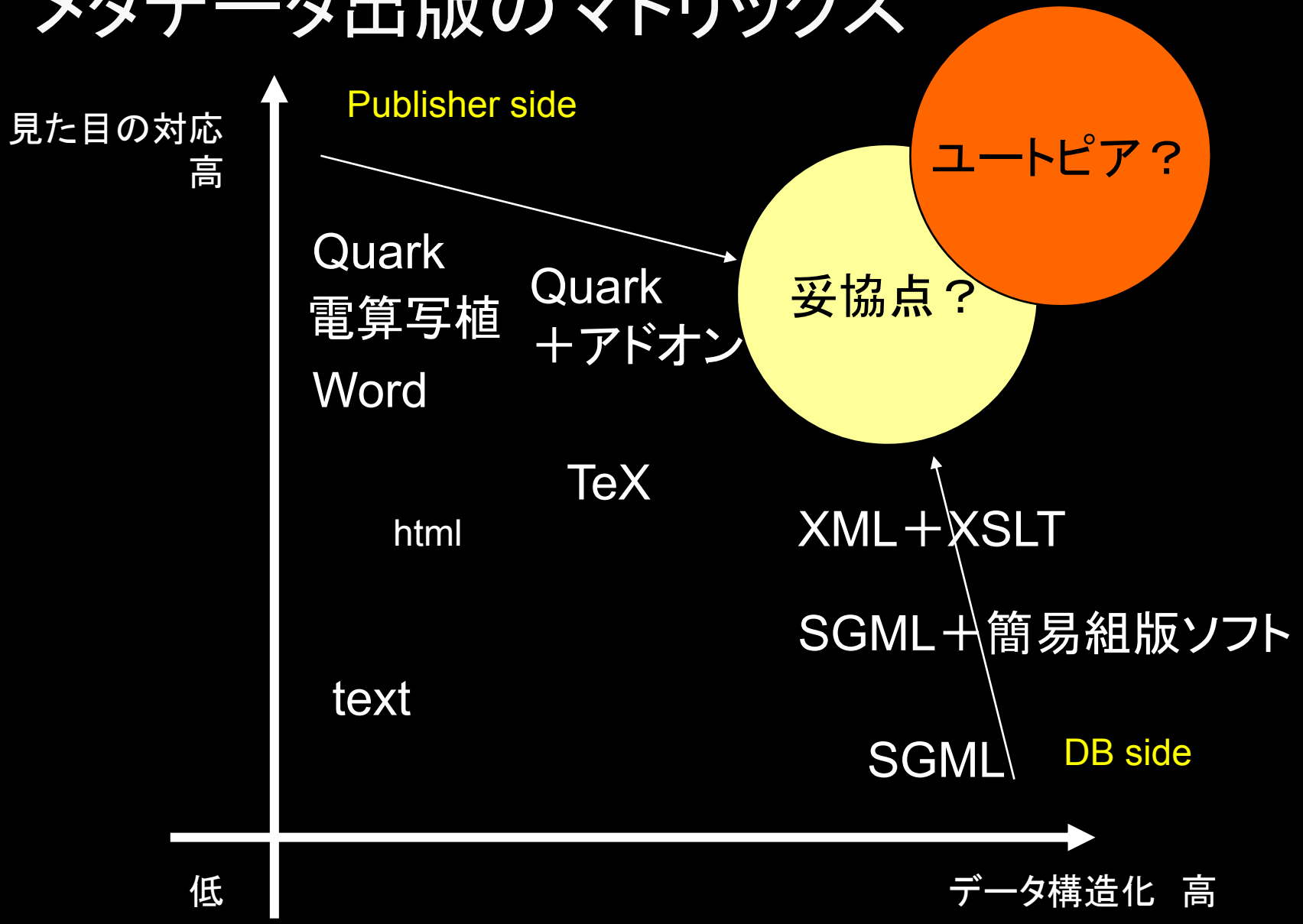
メタデータ(XML)出版のキモ

- 設計:どの程度タグをつけるべきか、つけたタグは最終的にどのように利用されるか
- 運用:誰が、どこで、いつ、タグをつけるか

Tips:

- タグが多いものから少ないものの変換は比較的容易だが、逆は難しい(面倒)
- タグ付けにおいて著者は信用できない(不特定多数の思想のばらつき)

メタデータ出版のマトリックス



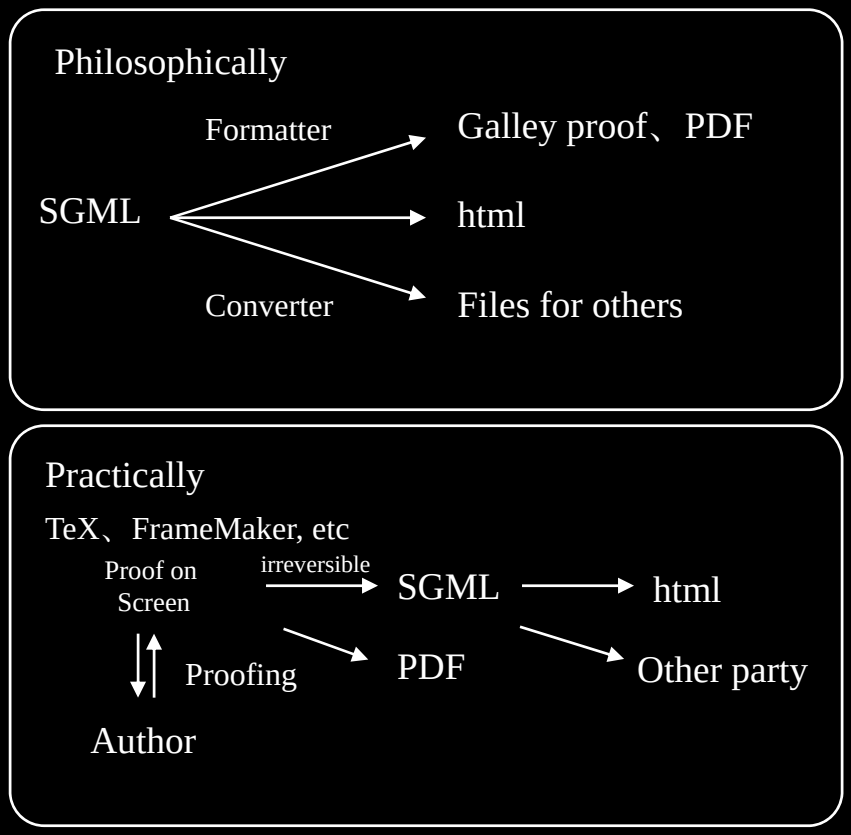


Fig.1　SGML出版の理想と現実

化学会では1989年からSGML出版に取り組んだが、2001年で一旦終了し、IPAPさんと同等のシステムに移行した。

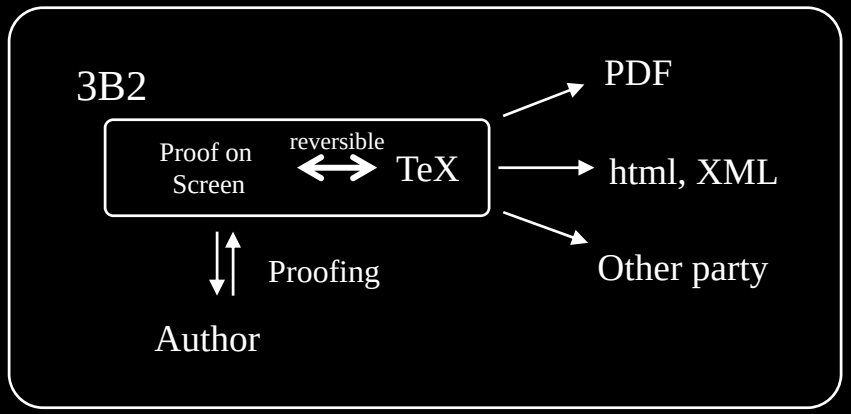


Fig.2　TeX-3B2 System

林、情報の科学と技術2002

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

本当に良くある（あった）質問（化学系）

- TeX-3B2は面白い
- なぜTeXじゃないといけないか？
- Wordじゃだめなのか？（著者Word率は2001年の時点で95%以上）

- だめではないがWordからTeXに変換するのは易しくはない。
- 数式が比較的少ない生物・医学あるいは応用分野系でも同様（TeX文化がなじみにくい）

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

Word-(Html)-TeX-3B2システム（WHT3)

- Wordから初期TeXまでは労働集約的処理
- 最初はテキストに落としてからコマンド（タグ）を直接打っていた
- 途中からhtmlコンバートで多少のスタイルからTeXへの変換処理を可能に
- いずれにせよ人件費の安い労働を利用

- 化学会としてはメタデータ出版を確立しつつ、版下作成コストを大幅に削減した（林、情報管理2003)
- 非常に現実的なシステム　XML出力も可能

Wordから直接XMLを出力できるか

- SGML出版も理念としては正しかったので、SGMLの欠点を補ったとされるXMLの利用については2002年以降も継続調査

- 一つの解決案としてeXtyles（2004登場）
- 当時は前述WHT3のシステムが安定して間もない頃のために、存在を認識しながらあくまで様子を見ていた。
- NLM-DTDの登場と浸透
- PorticoプロジェクトによるNLM-DTDのデファクトスタンダード化
- 組版システムのバージョンアップによるXMLの親和性アップ
- 再検討

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

eXtylesとは

- Inera社開発のMS-Wordのプラグインソフト
- 任意の著者ファイルから始めて
  - 不要なデータの削除
  - パラグラフ単位でのスタイル付けとタグ付け
  - パターンマッチングによる半自動編集
  - リファレンスの解析とタグ付け
  - リファレンスチェック
  - 図表引用、リファレンス番号の整合性チェック

→最終的なメタデータ（要素とコンテンツ作成）生成に必要な工程をWord内で行い→NLM-DTD　XML

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

eXtyles導入のメリット

- （ある時点での）NLM-DTDに基づいた質の高いXMLの出力が可能（Inera社CEOのRosenblum氏はNLM-DTDの共著者）←一からDTDに合わせてXMLを作成するのは大変
- 手入力では非常に非効率なXMLタグ付けが容易に（閉じタグの入力の手間　TeX”}”　XML”</要素名>）
- 初期XML出力までが早い（図表が少なければ著者Wordから数時間以内に出力することも十分に可能）→超早期公開
- 全文データ（html、xhtml）作成が容易
- ヒューマンリソースの確保がしやすい（Word処理）
- 学会編集側で処理できる可能性もある
- 赤入れ担当者の負担軽減（なくなりはしない）
- リファレンスリンクの多重チェック（PubMed, CrossRef)

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

eXtyles導入のデメリット

- 著者のWord利用が多くないと力を発揮できない
- 1OSの1ツールに依存することによるリスク

→著者や業界の動向は常にウォッチ

- 著者原稿の当たり外れによる前処理の負荷
- 労働集約的な作業にしにくい（今のところ）
- カスタマイズに一時的に負荷がかかる（便利にしたいなら）
- 導入コストの償却

→これはどのようなツール導入でも同じ

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

コンピュータリソースの変化

- 昔に比べてPCの性能が格段に向上
- ツール類の充実
- 従来面倒でかつマシンパワー＆マンパワーを要していた作業が手軽に＝TeXに比較すると煩雑ともいえるXMLのタグ付けも可能に
- また、そもそもユーザーがタグ付けを意識しないインターフェース作りが可能に

→ようやくXMLを実際的に効率よく作成する素地ができたと言えるか？

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

Word（→eXtyles）→XML→3B2

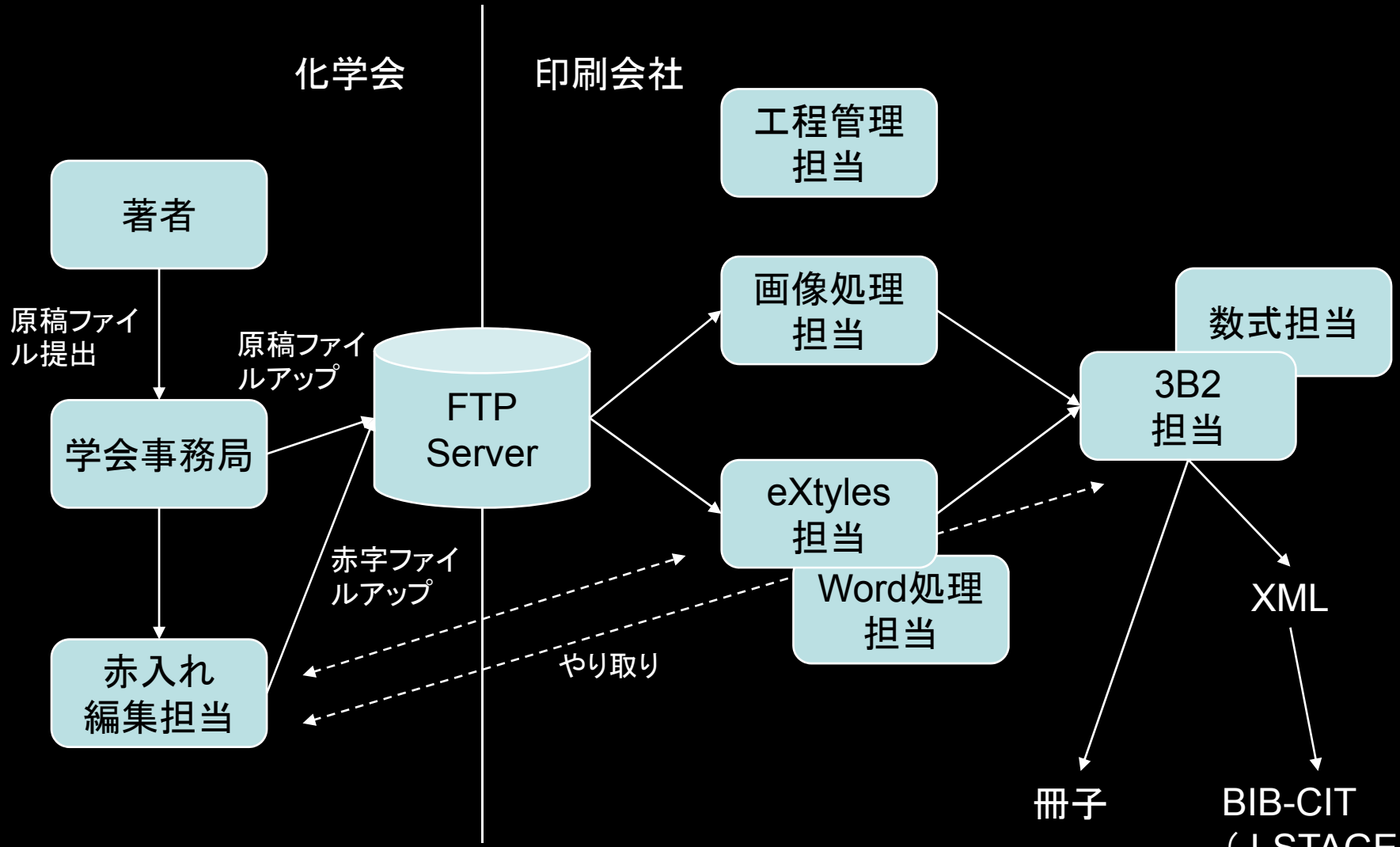
- MS-Word→eXtyles→XML→3B2　システムのオペレーション確立　(WeX3)

- 新しい赤入れと入稿手法
- 新しいメタデータ生成手法
- 新しい画像と数式の処理
- 新しい組版の手法
- 新しいメタデータ出力手法

→を検討（一部はほとんど変更なし）

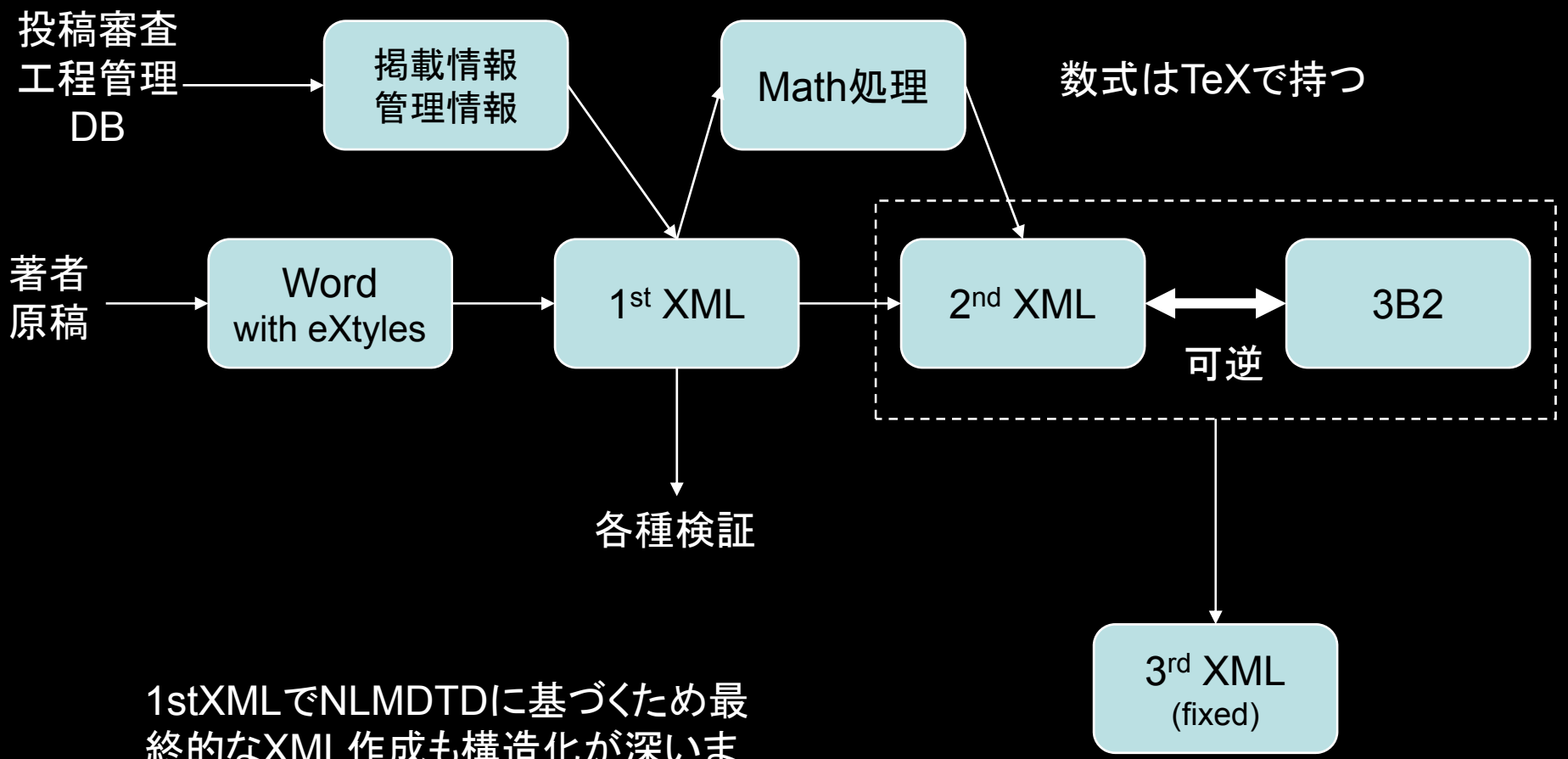
- 新しい役割分担とワークフローの検討

各オペレータの位置づけ



SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

XML作成とMath処理（簡易版）



1stXMLでNLMDTDに基づくため最終的なXML作成も構造化が深いまま（→Tips）

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

課題と展望

- XMLベースで進めることがどこまで有効か見極める（XMLありきにならないように）
- 「欧米に比較して」コスト対効果をどこまで出せるか（事業効率）
- 全文データ(xhtml）の実際的な利用

- NLM-DTD　XMLの国内での浸透具合を見る（人柱）
- 生物、医学系ジャーナルへの展開
- 和文誌対応
- 投稿、査読システムなどとの本格的な連携
  - Editorial Managerの例
  - J-STAGE

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

Web-oriented publishing

- Webによる研究者コミュニケーションの多様化
- 論文として書く前からのタグ入れの可能性（blog)
- 発行した後のコミュニケーション
- PCリソースとツールの発達および研究者の行動習慣の変化が、審査済みの情報からのlegacy publishingではなく研究情報が生まれる段階からのpublishingを加速する可能性が高い
- ただし、研究評価も変わらないといけないが（これが重要）
- いずれにせよ新しい時代の情報流通ソリューションが構築される可能性は十分にあり、そのときの標準（流通）メタデータの有力候補はXMLであろう

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

ただし、

- いつその時代が来るかはわからない。
- 事業運営、ビジネスの点からは現実的な手法とが必要。
- また、開発と安定運用のバランスをとることが望ましい。
- XMLを作るためにXMLを作ることは避けたい。

SPARC JAPAN
SPARC : Scholarly Publishing and Academic Resources Coalition

ご清聴ありがとうございました。