物質・材料研究におけるテキストデータマイニング：
効率的文献データ利活用

国立研究開発法人　物質・材料研究機構
石井　真史

革新的材料開発力強化（M³）プログラムを構成するMaterial Research Bank
材料R&Dデータの入り口（収集）から出口（サービス）まで
データ収集 → 処理 → 解析 → 蓄積 → 管理 → 公開
→ 駆動 → 協調的利用 → 競合的活用

- ❖ 情報の爆発的増加
- ❖ 学術論文を出典とするデータベースの維持・運用の進め方
- ❖ 機械学習や人工知能を駆使した効率化

Materials Research Bank (MRB)
– 材料データプラットフォームセンターのミッション–

先進データベース

- 文献からのデータマイニング
- 機械学習手法を取り入れた先進的データベース

データ収集と高付加価値化

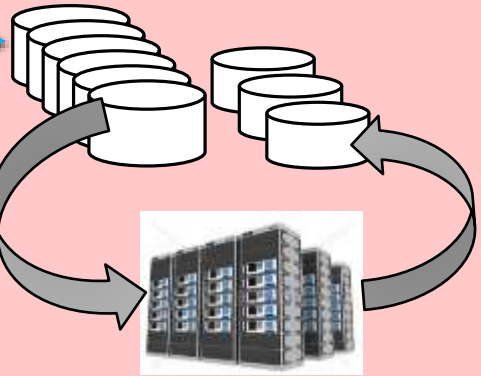
- 実験装置、計測機器、IoT
- 日本の大型研究施設から
- メタ情報とトレーサビリティを担保
- 信頼性のあるデータ、標準データ

NanotechJapan

SIP-MI統合システムによる利用技術の活用

材料情報統合データプラットフォーム

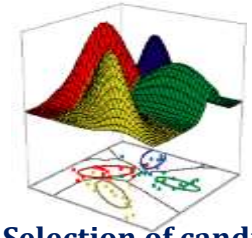
物質・材料データベース



解析サーバ

ツールやアプリケーション

連携・共同研究

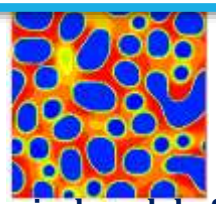


Selection of candidate materials

協調利用→競争活用

Visualization of materials

公開＋利用



Physical model of micro-structure

民間利用

学術利用

データ駆動型データパブリッシュ

NIMSの保有する材料基礎物性データベース群
学術論文からのデータ抽出

http://mits.nims.go.jp/

NIMS MatNavi

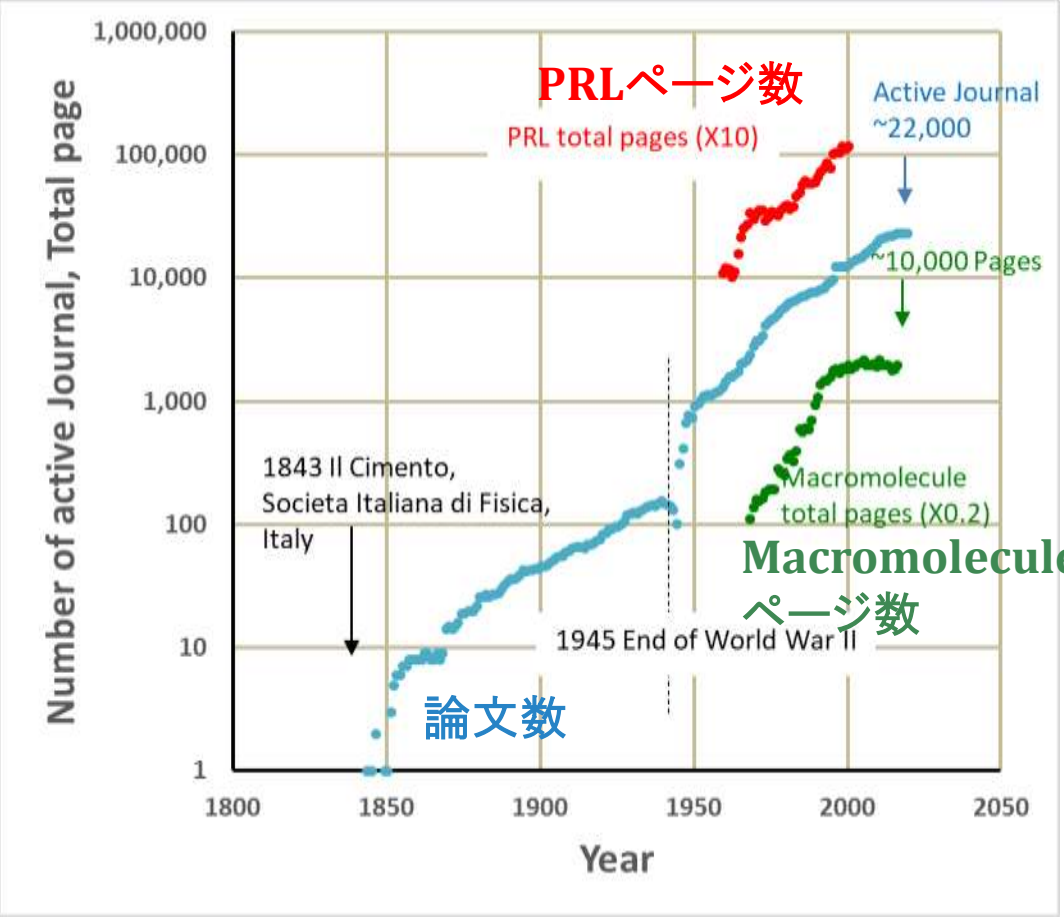
名称	内容	データ数
PoLyInfo	高分子	299,258
AtomWork	無機材料	55,000
Kakusan	拡散係数	4,242
SuperCon	超電導材料	32,791

学術論文からデータを人手で抽出

NIMSが誇る材料データベース

材料開発、教育、マテリアルズ・インフォマティクスでの利用

テキストデータマイニングの必要性
科学技術における情報量の増大



雑誌数=Scopus掲載数で近似

1843 Il Cimento (Italy)以来、学術雑誌数の指数関数的増加

第二次大戦以降の雑誌数の増加

10倍/50年

論文のページ数の増加

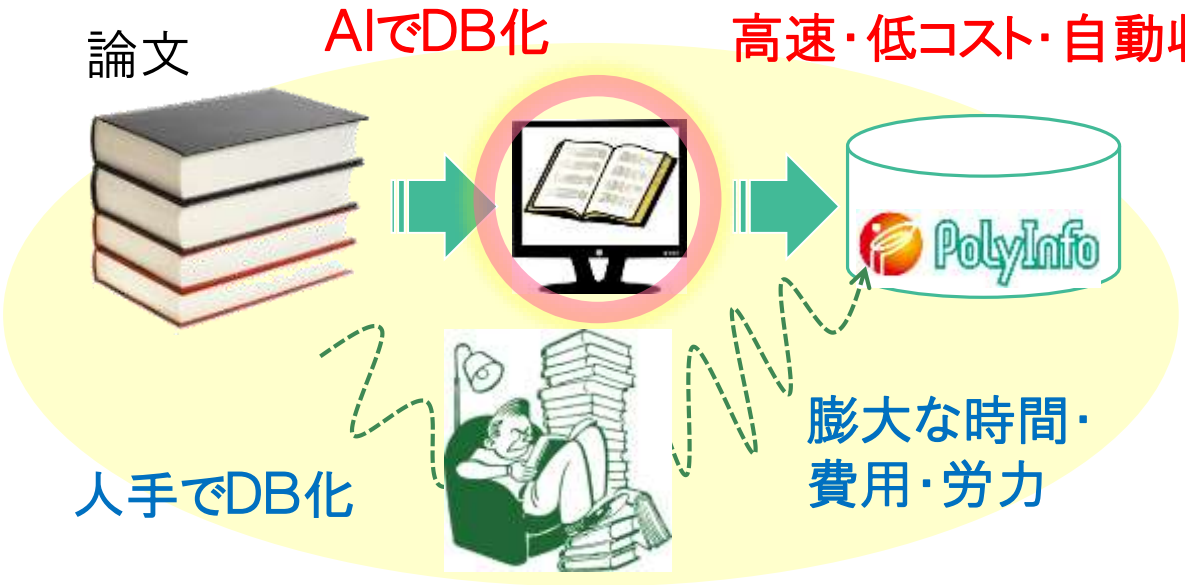
×10倍/50年

トータル

=100倍/50年

人手によるデータベース構築の限界

テキストデータマイニングによる論文からの自動データ抽出



201個の”polystyrene”試料に
46の違う名前。でも

みんな
”Polystyrene”

テクニカルな課題

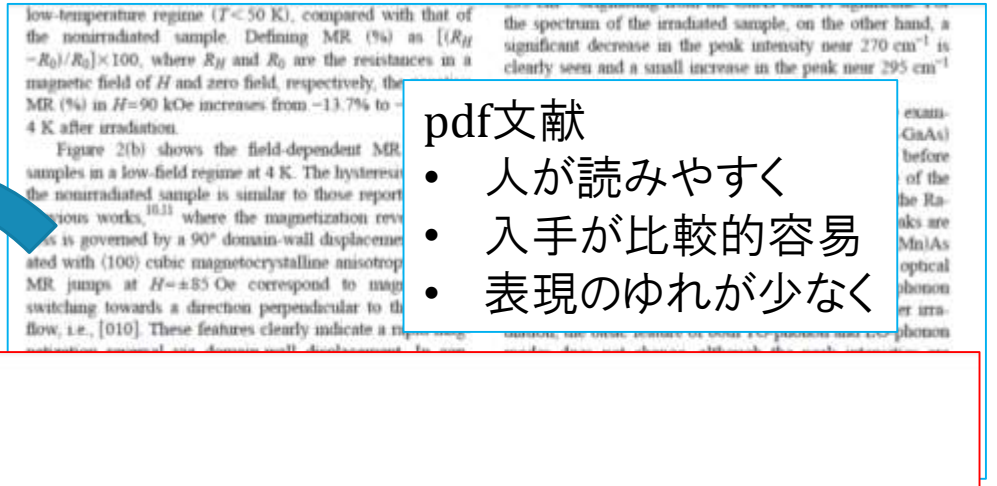
- 「物質名」の表記ゆれ
- 「特性名」の表記ゆれ
- 「単位」の表記ゆれ
- 「物性名」-「物性名」-物性値-「単位」の関係づけ

alpha,omega-PS-(Rf)2,Rf	omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.5
com. H-PS	omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.6
D1	omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.7
D2	omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.8
D3, D4 and D5	omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.9
fluoroalkylsilyl	polystyrene
Home H-PS	protiopolystyrene (h-PS)
hPS	PS
L18	PS (hPS)
L27	PS latex (micro-PS-SDS, PS-SDS-alpha, micro-PS-OP, conv-PS and r-PS)
L9	PS/washed carbon fiber composite(CH2Cl2 soln. impregnation)
linear polystyrene	PS/washed carbon fiber composite(cyclohexane soln. impregnation)
Linear PS	PS/washed carbon fiber composite(molten PS impregnation)
omega-lithium sulfonatopolystyrene (PSsul)	PS-111
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.3)]	PS1-Br
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.4)]	PS-2.5M
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.5)]	PS2-Br
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.6)]	PS-37
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.7)]	PS-600K
omega-lithium [illegible]topolystyrene//trimethyl[illegible]arylammonium chloride [PSsul-TMSAC(0.8)]	PS-950
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC(0.9)]	PS-Br
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.3	PS-H
omega-lithium sulfonatopolystyrene//trimethylstearylammonium chloride [PSsul-TMSAC];0.4	PSt

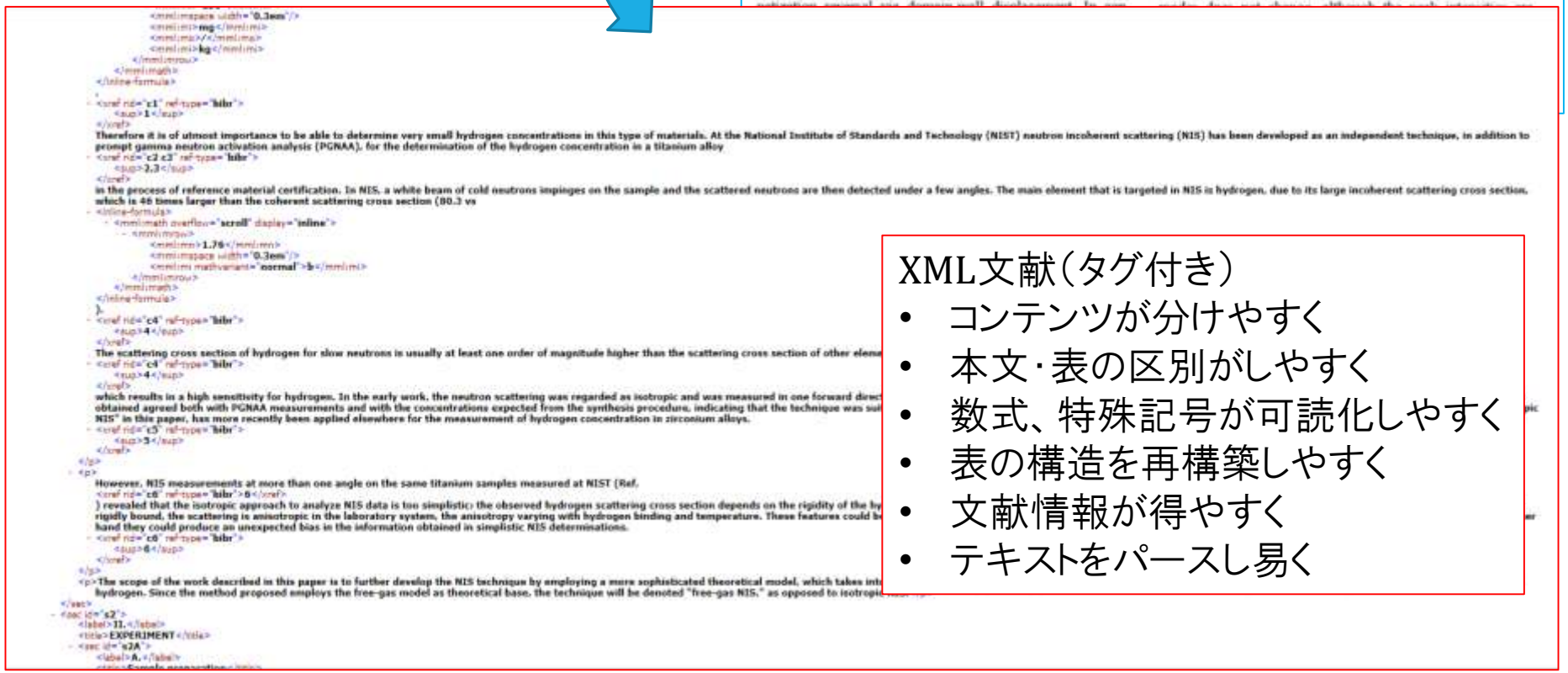
テクニカルな課題

・人が読易い文献 ≠ 機械が読易い論文

（機械が読みやすい）
文献リソースの確保

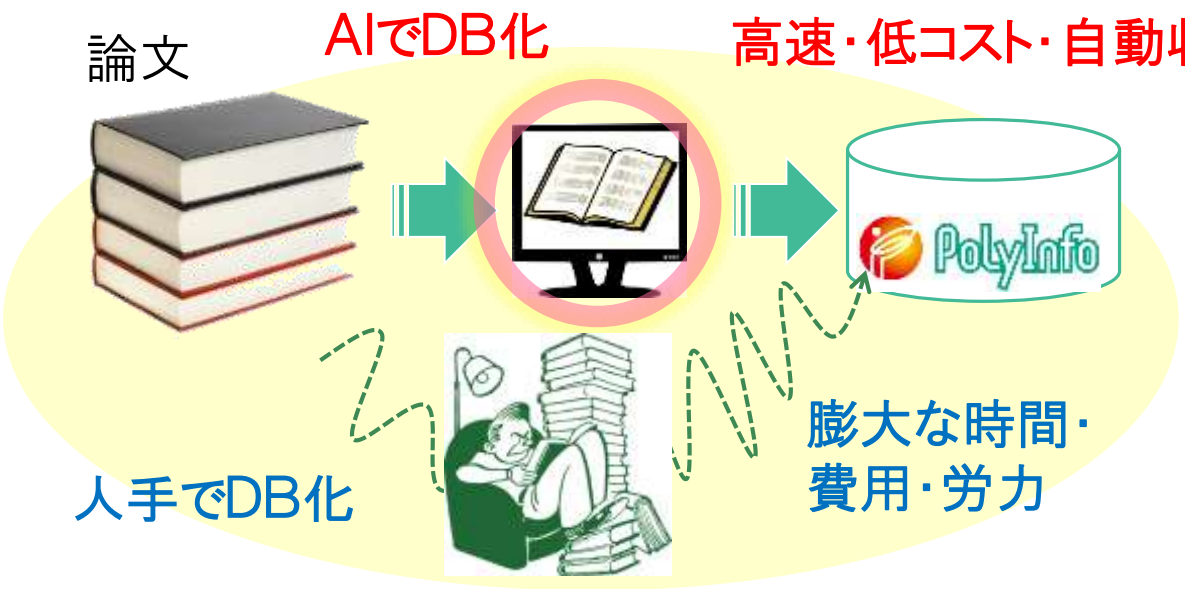


pdf文献
- 人が読みやすく
- 入手が比較的容易
- 表現のゆれが少なく

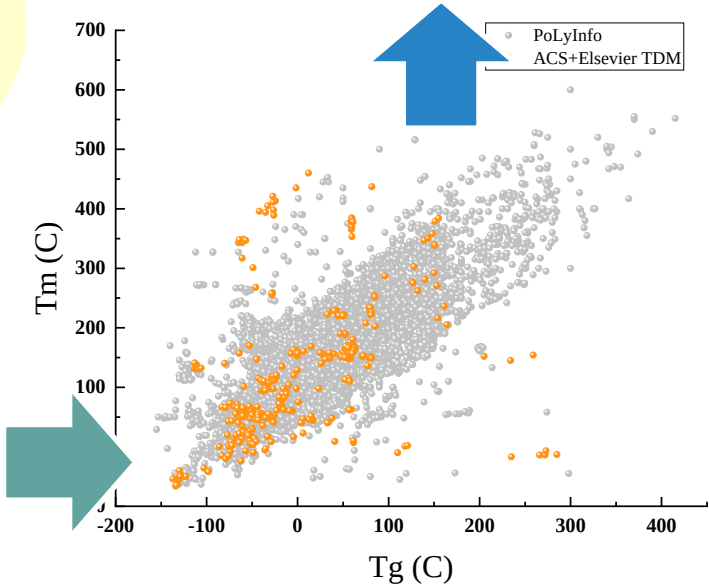


XML文献（タグ付き）
- コンテンツが分けやすく
- 本文・表の区別がしやすく
- 数式、特殊記号が可読化しやすく
- 表の構造を再構築しやすく
- 文献情報が得やすく
- テキストをパースし易く

テキストデータマイニングによる論文からの自動データ抽出



マテリアルズ・
インフォマティクスの
ためのデータ源



自動抽出成功例:

- 平均分子量X点
- 結晶化度X点
- 熱分解温度X点
- 溶融熱X点
- ・・・・他

X論文から
・X点のガラス転移温度
・X点の融点
の自動抽出に成功！

共同研究機関

理化学研究所 RIKEN

国立大学法人
奈良先端科学技術大学院大学
NARA INSTITUTE of SCIENCE and TECHNOLOGY

まとめ

　情報爆発が起きている現代で、NIMSの材料データベースを維持発展させる方法

→AI＋自然言語処理による文献からの自動データ抽出

→データ蓄積中！

論文からの自動データ抽出のカギ

- 物質名・物性名などの表記ゆれを吸収する辞書あるいはAIの作成
- 人手による抽出済データの活用（単語集や教師データとして）
- 機械が読みやすいリソースの確保