データ構造化諮問委員会　答申

データ構造化諮問委員会委員長　朝倉清高

日本放射光学会会長　横山利彦殿

１．本委員会発足の経緯

前回の諮問委員会“リモート実験等諮問委員会”において、オープンデータ、データポリシー、データフォーマット,メタデータを放射光学会としてまとめるよう要望があった。また、SPring-8 のデータセンター、NIMS の MDR などとの連携を見据えつつ、放射光データを蓄積・利用の試みが進んでいる。そこで、横山会長より、諮問委員会の設立が要請された。
横山会長からの要望として、
・データ構造化・蓄積・公開等に関する施設・利用者の考え方の集約
・データ公開・構造化の目的の学会として提示
・手法ごとの構造化データのフォーマットの策定。
があり、新たに本委員会を立ち上げた。

２．委員会の名前

データ構造化諮問委員会

３．委員名簿（敬称略）

朝倉清高	北海道大学触媒科学研究所
高橋幸生	東北大学国際放射光イノベーション・スマート研究センター
矢代　航	東北大学国際放射光イノベーション・スマート研究センター
若林裕助	東北大学理学研究科
仁谷浩明	高エネルギー加速器研究機構物質構造科学研究所
山田悠介	高エネルギー加速器研究機構物質構造科学研究所
木村正雄	高エネルギー加速器研究機構物質構造科学研究所
五十嵐教之	高エネルギー加速器研究機構物質構造科学研究所
石井真史	物質・材料研究機構
西堀英治	筑波大学　数理物質系　物理学域
松井文彦	分子科学研究所極端紫外光研究施設　JSSRR 庶務幹事
岡島敏浩	あいちシンクロトロン光センター
田渕雅夫	名古屋大学
稲田康宏	立命館大学生命科学部

栗栖源嗣	大阪大学蛋白質研究所
小野寛太	大阪大学工学研究科
矢橋牧名	理研
初井宇記	理研　データ処理系開発チーム
熊坂　崇	JASRI 構造生物学推進室　JSSRR 渉外幹事
為則雄祐	JASRI 分光推進室
登野健介	JASRI 散乱・イメージング推進室
玉作賢治	JASRI 回折・散乱推進室
松本崇博	JASRI
木下豊彦	JASRI
藤森伸一	日本原子力研究開発機構　原子力科学研究部門
松尾光一	広島大学放射光科学研究センター
廣沢一郎	九州シンクロトロン光研究センター
米山明男	九州シンクロトロン光研究センター
東純　平	佐賀大学　シンクロトロン光応用研究センター
原野貴幸	日本製鉄/日鉄ケミカル＆マテリアル
横山利彦	陪席　分子科学研究所　JSSRR 会長

４．委員会議事録（抄録）

これまでに合計 6 回の委員会をひらいた。

第 1 回　委員会　2022 年 4 月 5 日 10 時～　オンライン

今後の進め方について検討

第 2 回　委員会　2022 年 5 月 19 日　10 時から 12 時

初井宇記先生の講演　　理化学研究所「SPring-8 データセンター構想 データ構造化について」

大型データの蓄積受け渡しの問題、「データ創出・実験中のデータ処理・一時保管を担う」のが目的,メタデータの利用目的には、施設運営のため、ユーザーへのデータ一時保管機能提供

栗栖源嗣先生の講演　生体高分子構造データの現状

XRD のほか、核磁気共鳴、電子顕微鏡も含む、日米欧でデータを共有、共同組織を同じ編集方針、国際分業、一つのデータベースを作る。　PDBx/mmCIF フォーマット、生データではない。

人手と予算の問題、実験データの解析精度と理論との一致度を評価

全体会議で、班分けしてメタデータの検討することとなった。

第 3 回　委員会　2022 年 6 月 13 日 12 時～14 時

重藤知夫先生　　産総研「共通データフォーマットと独立可用性」

再現可能なメタデータを含むことにより、独立可用性を持つデータとなり、機械学習に適する。MaiML が JIS 規格として作られつつある。

石井真史先生　　NIMS「MDR XAFS DB　メタデータとデータ連携の取り組み」

NIMS MDR に XAFS DB ができ、施設を超えたデータの収集を開始した。

階層メタデータモデル,材料に一意の ID をつける。NIMS 材料辞書

全体会議

3 班に分けて検討

・分光スペクトル I(E)　XAFS など

・イメージング I(xy) z もあるかも。

・回折ディフラクション I(theta, phi, E …) 小角散乱も入る。

第 4 回　委員会　2022 年 10 月 14 日 13 時～15 時

３分科会（イメージング・スペクトル・回折）からの報告

矢代　航委員　イメージング班

SPring-8 ですでに整備した key-value

問題点　どのデータをのこすか（ダークファイル、中間ファイルなど）、試料回り環境情報・解析ソフトの情報の取り扱い、機器による出力フォーマットの違い

提案　日本全体での議論　データの独立可用性の確保（第 3 者による再解析が可能）電子カルテ

為則雄祐委員　スペクトル班

XAFS で検討　－NIMS DB にスペクトルの公開がすすんでいる。

手がかかるのは重複の key の整理。

第一階層　key の利用状況を調べた。

XAFS DB の第一階層（data info, facility file instrument measurement sample）

第一階層　施設や手法を横断して共通化

第二階層　施設依存

第三階層　ユーザー依存

検討事項

　エネルギー校正情報、標準データの定義、質の判定、実験ノート情報の電子可読化

自動記録と実験ノート記録

五十嵐教之委員　回折班

回折　IUCr 主導の CIF(crystallographic information file)を元にする。
問題点
測定手法・解析手段の記述
RIXS の取り扱いスペクトルにまたがる。

第 5 回目　委員会　2023 年 2 月 27 日（月）14:00-16:00　オンライン

XAFS メタデータの紹介　　YAML で記述し、Schema は GitHub で公開、物質辞書（MatVoc）の活用
為則、木村、松本、石井各委員より
為則委員：
初井宇記　委員　自動化 CT ビームライン BL28B2 でのスマート化の現状とデータ構造化
データ取得、解析、キュレーション、格納、利用
RDF（Resource Description Framework） 形式と Key Value　　RDF が正式である。　外部規約を明示できる　/　Key Value 簡易型

第 6 回目　委員会　2023 年 3 月 22 日（月）　16:00-17:00　オンライン

初井委員による RDF を中心とした説明
研究データ（頻繁な規約変更、捏造、共同研究者間の共有、施設との共有）と公開データ（安定したデータ、DOI、不特定多数の共有）の区別
RDF から Key Value へは変換可能　その逆はデータの内容の詳細の指示書が必要。

5. 答申

以上の議論と経緯を踏まえて以下を答申する。

5.1　データ蓄積と公開の意義と基本的なポリシー

データは科学・技術におけるもっとも基本的な成果であり、人類共通の財産である。データを正しく収集し、蓄積し、一定のルールに従い公開・共有することで、科学者・技術者の基本的な社会貢献となる。特に、放射光は、巨大施設を用いるビックサイエンスであり、社会からの多大な支援を受けている。したがって、放射光で得られたすべてのデータの維持・管理・公開は社会への放射光科学・技術の還元として重要である。一方で、秘匿されなければならないデータがある。個人や企業の損益に直接・間接的に関わるデータ、国民の生命や国家の安全に関わるデータなどである。こうした秘匿を必要とするデータについては、一定のルールを設けて、保護と公開をバランスよく実現することが求められる。ルールについては、社会全体の動向をみながら、今後慎重に策定する必要がある。

5.2　データの構造化の意義

上で述べた公開し、そのデータを利活用するためには、データを特徴付け、他のデータとの関連をつけることが必要であり、メタデータを含めて、一定のルールでデータの収集と蓄積がならされること必要である。また、データの公開や公共の利活用に当たっては、メタデータは、共通のルールに従い作成されることが重要である。これは、データを取得した個人あるいは小集団内においても将来、再利用する際にも重要な要件となる。今後さらに発展する機械学習によるデータの利用にはおいて、よく整ったメタデータとデータは必要不可欠である。

5.3　データの構造化における問題点

データの構造化では、データだけでなく、データに関連した様々な情報、データを取得した条件などのメタデータが必要になる。データの独立可用性が重要な概念となる。メタデータに関しては、共通のルールに則って作成される必要がある。一方データそのものについては、フォーマットの共通化が必要となる。こうした共通化をすることが最も重要でかつ困難な点である。

5.4　本委員会の役割と放射光学会への要望

(a)　本委員会は 5.5 に示すメタデータに関する提言と放射光学会への要望を行う。

(b)　本委員会は放射光学会に対し、データに関する常置委員会の設立を要請する。新たに設立された委員会において 5.6 に示したようなデータ公開へのルールづくり、データのフォーマットなどに議論を行う。

5.5 メタデータに関して提言と放射光学会への要望

5.5.1　メタデータのカテゴリー分け

メタデータは３つカテゴリー。

(ア) 第１カテゴリー　放射光実験全体に関する事項

(イ) 第２カテゴリー　3 大領域である分光法、回折法、イメージング法ごとに関わる共通の事項

(ウ) 第３カテゴリー　個々の手法ごとに特有の事項

にわけることができる。

5.5.2　データは公開データと研究データ

（１）研究データは、グループ内利用であり、変化が容易に起こると予想される。また、非構造メタデータも存在すると考えられる。

（２）公開データでは、メタデータが不特定多数の人に一義的に解釈されないといけない。公開データの場合には RDF（Resource description framework）などにすることで、安定的に情報を提供できることがのぞましい。Key Value 形式であっても RDF に変換できるようにすることで、独立可用性を担保できるようにすることが重要である。

5.5.3　第 1 カテゴリーメタデータについての提言

第 1 カテゴリーに属する必須メタデータについては表 1 にまとめた。ここでは、Key-Value で示してあるが、いずれは、RDF に返還されるなどの工夫が必要と考える。

5.5.4　サンプル名、物質名、DOI

サンプル名、物質名はデータを紐付けるのに必須のメタデータである。その定義、名前の統一については、放射光以外のコミュニティーを巻き込んだ広範な議論が必要である。サンプル名と物質名は区別を付けることが必要である。これは区別して用いないと、混乱の元になる。

サンプル名(NameOFSample)は測定したサンプル固有の名前であり、標準サンプル、未知サンプルを含め全てのデータについている必要がある。この名前はローカルな名前でよいが 1:1 で物理的実態を持つサンプルと直接結びついている必要がある。これにより、他の手法による測定データと一体となり、サンプルのキャラクタリゼーションを行うことができる。またサンプル名がユニバーサルな名前で登録できるようになることで、公共の利用が促進される。例としては、各研究者が付けている通しサンプル番号、市販薬品の Lot 番号などである。

一方、物質名（NameOFSubstance）は、サンプルが属する物質に関する名前である。標準サンプルでは、測定前に決定され,記録されていることが必要となる。また物質名は統一

したルールにしたがって、書かれなければならない。(たとえば、α酸化アルミニウムとかαAl2O3 など)名寄せのための辞書の作成（NIMS の MatVoc など）も必要になる。CAS 番号、化学記号などもその一つになり得る。合金のように組成をあわせて書いておく必要もある。こうした、連続的に変化するもの、不均一なものについては、多くの手法を総合してキャラクタリゼーションをする必要があり、前述した測定物質と 1:1 に対応したサンプル名を付けることが重要になる。未知試料については、この物質名は測定が終わってから決まることが多い。

最後に NIMS の MDR（Material Data Repository:mdr.nims.go.jp）にあるように DOI（Digital Object Identifier）　が各データに付されることが望ましい。これにより、データを特定できるとともに、引用を通じてデータ蓄積とデータ公開のインセンティブとなる。

5.5.5　第 2 カテゴリー、第 3 カテゴリーに属するメタデータについて

第 2，第 3 カテゴリーを区別することは難しい。また、第 3 カテゴリーのメタデータは個別手法に特化したものであり、本委員会が決めることは難しく、各手法のコミュニテニティ-に委ねるべきである。放射光学会は各手法のコミュニティーに積極的に働きかけ、データフォーマットの共通化と共通メタデータ作成を打診することを要望する。新しい常置委員会では、各コミュニティーから上がってきたメタデータのオーソライズとこれを元に第 2 カテゴリー作成を要望する。なお、第 3 カテゴリーの参考として、IUCr の CIF 形式(https://www.iucr.org/resources/cif)や日本 XAFS 研究会（https://www.jxafs.org/xafs-database/）が提出しているメタデータ、今後公開されると言われているMaiML などがあげられる。

一方、放射光学会は各施設に対して、ビームラインや装置の維持運営に必要なメタデータとその定義の提出を依頼することを要望する。これをもとにして、新しい常置委員会で、共通メタデータを策定する。これによりつくられたデータベースは各施設・各装置の維持管理に利用するだけでなく、施設間の協力関係を一層発展できることになると期待される。

5.5.6　データベースを維持し利用を促進する活動の推進に関する要望

データを共有し、再利用する考えを学会として基調とすること、データを共有する基盤（データベース）の維持を支援し、運用に対する寄与と提言の両輪を持つことを要望する。材料研究のデータは、バイオ関係に比べると、共有・再利用に対する意識がはるかに低く、アカデミアであってもメタデータの付与やリポジトリ・データベースへのオリジナルデータの登録はほとんど行われていない。放射光学会は、社会への放射光科学・技術の還元を先導する立場を明らかにし、データの共有を標準とする意識変革の活動を進めることを提言する。実際にデータを受け入れる共有基盤（データベース）における、クオリティを含むデータマネージメントを検討し、実際にデータを循環させる機関に、運用に関する要望を効果的に伝えること、そのためにメタデータのみならずデータベース自体の構

築・維持に寄与することなど、データ共有の全体方針の立案と実施を提言する。

5.6　次期に設置を期待するデータベース常置委員会における想定されるミッション

1. データ公開のルールづくりと普及
2. 第 2，第 3 カテゴリーのメタデータのオーソライズと策定および普及
3. 個別手法ごとのデータフォーマットのオーソライズと普及
4. サンプル名のユニバーサルルール、物質名の統一と辞書の策定および普及
5. 公開データフォーマットの検討、公開用データのメタデータの確立
6. 他のコミュニティーとの連絡、他国との連携
7. 今後公開が予定されている MaiML との整合性の確保

第 1 カテゴリーにおける必須のメタデータ

Facility Name	Facility Name	String	SPring8, SACLA, NewSUBARU, PF, PF-AR, Aichi SR, SAGA-LS, UVSOR, Rits-SR, HiSOR, NanoTerasu, INS-SOR
BeamlineANDBranch	ビームライン	String	BL36XU, BL10B, NW10A
SRElectronEnergy	測定時の蓄積エネルギー	Numerical	2.5
UnitOFSRElectronEnergy	上記単位	SRring(GeV,MeV)	GeV
SRCurrent	測定開始時の電流値	Numerical	450
UnitOFSRCurrent	上記単位	String(mA)	mA
Polarization	偏光方向、特性	String	Linear（H）,Linear(45*1) Circular(+)
Method	測定法	String	XAFS
DataDimenstion	データの次元	Numerical	2（スペクトル、粉末 X 線回折） 3（イメージング）
Axis1	次元 1 の定義	String	Energy, 2q

UnitOFAxis1	上記単位	String	eV, °
PhyscialMeangingOFAxis1	上記の意味	String	入射 X 線のエネルギー
Axis 2	次元 2 の定義	String	吸収係数、強度
UnitOFAxis2	上記単位	String	None, cps
PhysicalMeaningOFAxis2	上記の意味	String	
AxisN			
UnitOFAxis			
PhysicalMeaningOFAxisN			
NameOFSample	サンプル固有の名前	Str	0000-0003-1077-5996_202008011000JST_Al2O3（ORCID_（取得年月日時間）_物質名） グローバル、ローカルであっても一義的に物理的な名前のとくていができる。
NameOFSubstance	所属する物質群名	Str	グラファイト, ダイアモンドなど測定物質が属する物質群 MatVoc などによる名寄せが必要
DOI	Digital Object Identifier	Str	https://doi.org/10.48505/nims.3344