

第3回 SPARC Japan セミナー2016

「科学的知識創成の新たな標準基盤へ向けて：オープンサイエンス再考」

材料科学分野における データ利用のライセンスの考え方 - 図書館からの視点

小野寺 千栄

(物質・材料研究機構)

講演要旨



データには著作権がないため、利用に際してはライセンスが重要となる。本発表では、物質・材料研究機構で2016年に開催した3回の連続セミナーをもとに、データ利用におけるライセンスの考え方について整理し、データ利用の実例について紹介する。また、図書館はそこでどのような役割を担うことができるのかを考える。



小野寺 千栄

物質・材料研究機構 図書館員。大学図書館等での勤務を経て、2015年より現職。研究情報の受発信を含めた図書館業務全般に携わっている。

今日は材料科学分野におけるデータ利用のライセンスの考え方について、図書館からの視点でご紹介します。まず、物質・材料研究機構（NIMS）についてご紹介した後、NIMS で開催した「オープンサイエンスと著作権」のセミナーについて、次に NIMS においてデータがどのように扱われているのか具体的な事例をご紹介します、最後にデータ利用について図書館がどのように関わることができるのかを皆さんと一緒に考えられればと思っています。

1.NIMS の概要

NIMS は国内で唯一の物質・材料科学の専門機関で

す。2016年10月には理化学研究所、産業技術総合研究所と共に、特定国立研究開発法人に選定されました。組織のミッションとして、物質・材料科学技術に関する基礎研究および基盤的研究開発、研究開発成果の普及とその活用の促進、NIMS の施設および設備の共用、研究者・技術者の育成およびその資質の向上の四つを掲げています。職員数は約1,500人で、そのうちの約半数が研究に関わる職員です。私はそこで図書館司書として働いています。

図書館では、研究に関わる職員へのアドボカシー活動として、セミナーを開催しています。2016年は「オープンサイエンスと著作権」について3回の連続

セミナーを開催しました。NIMS のデータ利用についてご紹介する前に、データの著作権についての考え方を整理するため、まず、このセミナーの概要をご紹介します。

第1回は2016年1月に開催しました(図1)。講師は、内閣府の真子博さん、文科省の著作権分科会の委員でもある末吉互弁護士です。約100名の参加者のうち、NIMS 内部からは研究系職員約6%に当たる48名の参加がありました。真子さんにはオープンサイエンス政策の背景と今後の展開についてお話しいただき、末吉弁護士にはオープンサイエンスに必要な著作権、特に論文と研究データの著作権についてお話しいただきました。

講演では、(1) デジタル著作物を利用するには現行の著作権法は不十分であり、足りないところは契約で補うことができる、(2) 著作物とは、思想または感情を創作的に表現したものであり、データは著作物には該当せず、著作権法の対象にはならない、(3) そのため、第三者がデータを利用する際の条件は、契約で規定することになる、とお話しいただきました。この第1回の反響を受けて、第2回のセミナーを開催しました。

第2回はデータ共有を切り口に開催しました(図2)。講師は出版社、研究者、弁護士の3名でした。出版社からは、ネイチャー・パブリッシング・グループが発行しているデータジャーナル「Scientific Data」のエディターVarsha Khodiyar さんにお越しいただきました。研究者としては、この後、NIMS の事例でもご紹介する高分子データベースのデータ収集について桑島

功さんに、弁護士として第1回に引き続き、末吉弁護士にお話しいただきました。

Varsha さんからは、データ共有のプラットフォームとしてのデータジャーナルについてお話しいただきました。末吉弁護士からはあらためて、データは著作物に当たらないこと、そのためデータを収集して提供するデータベースも著作物として認められることは過去の判例を見ても非常に稀であり、データの利用は、データ提供元と利用者との契約によって規定されるということをお話しいただきました。

最後となる第3回のセミナーでは、材料・科学分野においてもデータ駆動型研究が推進されている背景から、テキスト・データマイニングや人工知能といった次世代の知的財産権を取り上げ、弁護士と出版社から講師をお迎えしました(図3)。

まず、内閣府の次世代知財システム検討委員会の委員でもある福井健策弁護士からは、人工知能をめぐる知的財産の論点として、三つのポイントをお話しいただきました。一つ目は、人工知能の学習用データの権利を著作権で保護するかどうか、保護する場合はどこ

図書セミナー「オープンサイエンスと著作権」第1回

- 2016年1月8日開催
- テーマ：オープンサイエンスと著作権
- 講師：内閣府、弁護士
- 参加者：101名
 - ・NIMS (研究系職員)：48名
 - ・外部：53名 (研究機関、図書館、政府機関、出版社・学会、大学)
- 概要：
 - ・オープンサイエンス政策への取り組みと今後の展開。
 - ・論文と研究データの著作権。現行の著作権法はデジタル開示された著作物の利用には不十分であり、契約で補う。データは著作物ではなく、著作権法の対象にならないため、利用者が利用する条件は契約で規定する。

(図1)

図書セミナー「オープンサイエンスと著作権」第2回

- 2016年3月10日開催
- テーマ：(続) オープンサイエンスと著作権 - データ共有
- 講師：出版社、研究者、弁護士
- 参加者：83名
 - ・NIMS (研究系職員)：51名
 - ・外部：32名 (企業、研究機関、政府機関、大学、図書館、出版社)
- 概要：
 - ・データキュレーターから見るデータ共有 (データジャーナル)。
 - ・NIMSにおけるデータ収集と共有。
 - ・データベースが著作物として認められることは、過去の判例を見ても非常に稀。著作物ではないデータは自由に利用できるが、データ出典元との契約があれば、その契約に従う。

(図2)

図書セミナー「オープンサイエンスと著作権」第3回

- 2016年11月24日開催
- テーマ：(続々)オープンサイエンスと著作権 - データマイニングからAIへ
- 講師：弁護士、出版社
- 参加者：82名
 - ・NIMS (研究系職員)：61名
 - ・外部：21名 (研究機関、大学、企業、出版社、政府機関、図書館)
- 概要：
 - ・AIをめぐる知的財産権の論点について：AIの学習用データの権利をどこまで著作権で保護するか。アルゴリズム・学習モデルの保護範囲 (あるいは特許、プログラムの著作権として保護)、AIによる自動生成物は著作物か否か。
 - ・出版社におけるテキストデータマイニングの実例紹介。

(図3)

までを保護するのか。二つ目は、アルゴリズム・学習モデルの保護範囲をどうするか、特許やプログラムの著作物として保護するのか。三つ目は、人工知能による自動生成物を著作物とするかどうかということです。

出版社からは、エルゼビア社のヴァイスプレジデントの Anders Karlsson 氏にお越しいただき、エルゼビア社が提供している研究者向けのテキスト・データマイニングについて紹介していただきました。

3 回のセミナーを通じて、(1) 論文は著作物として著作権法で保護されるが、データは著作物ではないということ、(2) だからといって、何をやってもいいのではなく、契約によって条件を設定して公開できるということ、(3) 人工知能が絡んできて、知的財産権は今まさに議論が進んでいるところだ、と整理できました。

2.NIMS におけるデータ

データの著作権について整理できたところで、NIMS におけるデータ利用の実例を、作成者、利用者、提供者の三つの立場からご紹介します。

2-1.Case1

Case1 はデータの作成者としての関わりです (図 4)。研究データのメタデータ付与については以前から取り組まれていましたが、データ駆動型研究ではデータ共有がますます重要になるため、メタデータ作成の効率化・共有化が進められようとしています。

その一つが、計測データのメタデータ記録システムの開発です。NIMS 内の計測系研究グループで検討ワ

NIMSにおけるデータ：case 1 作成者として

- 研究データのメタデータ付与の効率化、共有化の取り組み
- 計測データのメタデータの記録システム

課題	対応方法
手書きのラボノート	データ採取時にデジタル化して記録
装置の操作で手がふさがる	音声入力
全てを入力するのは面倒	可能な箇所はシステムが自動計算
実験手順と入力項目の順番の相違	入力したい項目を入力したいタイミングで入力
装置が異なると、語句が異なる	語句の違いを吸収するためのテンプレート
メタデータと実験データの対応付け	メタデータ記録システムに実験データも保存

- 学会でも議論を共有して、課題解決に取り組む

(図 4)

ーキングが組織され、その議論から開発されたシステムです。研究者の立場からさまざまな工夫が凝らされています。

まず、計測データのメタデータを記録するときの課題として、実験時に手書きのラボノートしか記録していない場合、別途デジタル化作業が必要となります。そこで、計測装置を制御する PC からでも利用できるよう、ブラウザベースの入力インターフェースによって、データ採取時にデジタル化して記録できるようになっています。

また、装置の操作で手がふさがるため、音声入力にも対応していたり、全ての項目を入力しなくても済むように、可能な箇所はシステムが自動計算してくれるようになっています。

実験手順と入力項目の順番が違っていると画面を行ったり来たりしなければならぬので、入力したい項目を入力したいタイミングで入力できるようになっています。また、装置が異なると語句が異なるため、その違いを吸収するためのメタデータテンプレートが用意されています。そして、メタデータと実験データ (生データ) を確実に対応付けるため、メタデータ記録システムの中に実験データも保存できるようになっています。メタデータ付与については、組織内だけでなく、学会でも議論が共有され、課題解決のため分野全員での取り組みが進められています。

2-2.Case2

Case2 は、データの利用者としての関わりです (図 5)。NIMS では MatNavi というサイトで、材料開発に

NIMSにおけるデータ：case 2 利用者として

～ データサイエンティストによるデータ収集 ～

- 物質・材料データベースサイト (MatNavi) > 高分子データベース
- 高分子材料設計に必要とされるデータの提供
- データ収集元：出版論文 (購読誌、OA誌) から **データのみ** を収集
- 作業工程

文献管理

↓

データ収集

↓

物性データ収集

↓

詳細データ作成

↓

その他ファイル作成

↓

インターネット公開

データ収集・詳細データ作成はデータサイエンティストとして研究OBが担当

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

(図 5)

役立つさまざまなデータベースを公開しています。MatNavi はユーザー登録すれば無料で利用できます。今日は MatNavi で提供しているデータベースの中から、高分子データベース「PoLyInfo」のデータ収集についてご紹介します。

PoLyInfo は高分子設計に必要なデータを提供するデータベースです。データの収集元は、購読誌を含む出版論文で、その中からデータのみを収集しています。論文を収集元としている理由は、データの所在が明らかなこと、膨大な数が確実に出版されて公開されていること、必要とする情報が含まれていること等が挙げられます。

作業工程は、まず収集対象の文献を選定します。高分子分野は年に約 3,000 論文が出版されているのですが、スクリーニングによって 600~700 文献まで絞り込みます。次に、その文献から物性データを収集して、その物性データから構造検索などを可能にするための辞書データを作成します。その他の必要なファイルを準備して、高分子データベースとして公開しています。このうち、データ収集と辞書データ作成については、データサイエンティストとして、高分子研究 OB や有機化学研究 OB が担当しています。Case2 では、データサイエンティストによるデータ収集についてご紹介しました。

2-3.Case3

Case3 も利用者としての関わりです (図 6)。マテリアルキュレーションという手法によって、新しい材料探索を研究している方の例をご紹介します。それには

NIMSにおけるデータ：case 3 利用者として

～ソフトによるデータ収集～

- マテリアルキュレーションのためのデータ収集
- 背景：材料探索の前提となる多量のデータを取得する
- 方法：スキャンした論文PDFのグラフから、**数値データを抽出**するソフト開発
- 開発したソフトを特許取得し、希望者とライセンス契約
- ソフト自体はデータを持たない
- 著作権上の留意点については、研究者自身が弁護士に相談している (ソフトの配布は問題ないが、**第三者がソフトを利用してグラフから抽出したデータをグラフに描画することは、著作権侵害の責任を負う可能性がある**)

(図 6)

多くのデータを取得する必要がありますが、ボーンデジタルではない論文のデータも必要だったため、スキャンした論文 PDF のグラフから数値データを抽出するソフトを開発されました。同様の無料ソフトもあったそうですが、使い勝手が良くなく、効率的にデータ取得できなかったため、自分の研究用途に合うソフトを開発したところ、企業からも自社の古いデータをデジタル化したいというニーズがあり、特許を取得してライセンス契約することになったそうです。

ソフトを提供する際の著作権上の留意点については、研究者自身が弁護士に相談しており、グラフからデータを抽出することは著作権法上の問題はないが、抽出したデータをグラフに描画することは、著作権侵害の責任を負う可能性があるため、ソフト利用者にもそのように伝えているそうです。Case3 ではソフトによるデータ収集についてご紹介しました。

2-4.Case4

Case4 は提供者としての関わりです (図 7)。MatNavi で提供しているデータシートのうち、クリープデータシートがどのように提供されているかをご紹介します。

その前にクリープとは何か、ごく簡単に説明します。クリープとは、物体に一体の荷重を加え続けることでゆがみが生じる現象のことをいいます。NIMS では高温下で金属を引っ張り続けたゆがみを測定していて、長い試験では約 40 年引っ張り続けて、世界最長記録を更新しました。

このクリープデータを「無償」「有償」「詳細データ

NIMSにおけるデータ：case 4 提供者として(1)

～段階的な提供範囲の限定～

- クリープデータシート
- 公開方法
- 1. 無償
- 印刷物 (配布申請機関のみ)
- Web (要ユーザー登録、閲覧のみ、印刷・保存不可)
- 2. 有償 (英国 Grantak・デザイン社を通じて)
- データベースソフトウェアとのセット販売
- 3. 詳細データの個別契約 (NIMSとの直接契約)

(図 7)

の個別契約」という三つの方法で提供しています。無償提供は、印刷物とウェブの MatNavi です。印刷物は、配布希望の申請をした機関のみに配布していて、申請の際には利用目的も明記するようになっていました。これはデータを営利目的で利用されないように管理するため、過去には配布を拒否したケースもあったそうです。ウェブの場合はユーザー登録により、利用を管理しています。ウェブ利用においては、閲覧のみで印刷や保存ができないようになっていたため、無償提供の場合、デジタルデータは一切提供していません。

有償提供のデータは、無償提供しているものと基本的には同じですが、イギリスのグランタ・デザイン (Granta Design) 社が開発しているデータベースソフトウェアとセットで有償販売されています。さらに詳細なデータが欲しいというユーザーとは、提供内容を個別に契約しますが、これは NIMS との直接契約になります。

提供方法によって、利用を認める範囲は異なります (図 8)。まず、無償配布の場合、データをプロットしたり、データ解析結果を公表することは可能ですが、データを営利目的で利用することは認められていません。二次利用をコントロールする目的もあって、利用は登録制となっています。

有償販売の場合、その範囲はグランタ・デザイン社との契約に基づくことになります。詳細データを個別に契約する場合も、その範囲はケース・バイ・ケースですが、データそのものを開示しないこと、第三者へ提供しないこと、データを基に解析した結果は公開してもいいという3点は、原則として共通しています。

NIMSにおけるデータ：case 4 提供者として(2)

- 利用を認める範囲
- 1. 無償配布
 - OK：データをグラフにプロットする、データ解析結果を公表する
 - NG：データの営利目的利用
⇒ 配布/ユーザー申請時の理由によっては、拒否することもある
- 2. 有償販売
 - 利用範囲は契約に基づく
- 3. 詳細データの個別契約
 - データ開示禁止
 - 第三者への提供禁止
 - 解析結果の公表は可能

(図 8)

Case4 では、データ提供の範囲を段階的に限定している事例についてご紹介しました。

2-5. Case5

最後も提供者としてのケースです (図 9)。NIMS では情報統合型物質・材料開発イニシアティブ (MI²I) を進めています。これは JST のイノベーションハブ構築支援事業の採択課題の一つで、データ科学をこれまでの物質・材料科学に融合させることによって、新しい物質・材料科学の研究を加速する取り組みの場となることを目指しています。

活動の根幹を成すものの一つに、コンソーシアム活動があります。活動の概念図が右下の図です。データベースやツールを会員が使用することで、会員が持つコアな技術が拡大し、コンソーシアムで提供されるスクールや研究者のネットワークが技術の開発を促進します。技術が成長することで、雨粒のようにイニシアティブの中に落ちてきて、個別の共同研究を開始する契機になることを目指しています。

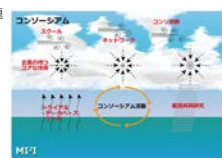
このため、研究コミュニティの発展に向けて、会員間でできる限りのデータ共有を目指しており、データ共有のためのプラットフォームを NIMS が提供しています。Case5 では、会員の利用規則という契約による、限定的なデータ提供の事例についてご紹介しました。

いずれのケースでも著作権法の範囲内とそうではないところを明確に区別して、著作権法の及ばない範囲については、契約でその利用範囲をコントロールしていることがお分かりいただけたかと思います。

NIMSにおけるデータ：case 5 提供者として

～ 限定的な提供範囲 ～

- 情報統合型物質・材料開発イニシアティブ (MI²I)
- JST のイノベーションハブ構築支援事業の採択課題
- データ駆動型の物質・材料研究への取り組み
- コンソーシアム会員を募って活動
- 会員の利用規則に沿ったデータ利用
- データプラットフォームは NIMS が提供



コンソーシアム活動の概念図
<http://www.nims.go.jp/MI-1/h6g4rf00000002oc-att/h6g4rf00000002pe.pdf>

(図 9)

3. 図書館業務とデータ

ここまで NIMS の研究者とデータとの関わりをご紹介しましたが、この中には図書館は関わっていません。データを扱うにはその分野の専門知識も必要ですが、図書館員がその必要な知識を習得するのを待っている時間的な余裕はありません。そこで、図書館員が今あるスキルでできることは何か、三つ挙げられるのではないかと思います(図 10)。

まず、図書館の伝統的業務である資料の保存と管理のスキルを生かして、研究者にデータの保存と管理に必要な情報を伝え、それを通じて図書館員もデータの扱いを学ぶということです。これについては、この後、尾城さんから詳しくご紹介いただけるかと思います。

次に、機関リポジトリのノウハウを生かして、研究者に保存と提供のプラットフォームを用意し、その対話の中で図書館員も必要なプラットフォームについて理解を深めるということです。NIMS では、文献をセルフアーカイブするリポジトリとは別に、マルチメディアデータをセルフアーカイブする imeji というリポジトリを研究者に提供しています。imeji ではメタデータ項目を研究者自身が設定できるので、拡張的なメタデータ管理ができるようになっていますが、先にご紹介した研究者自身が開発したメタデータ記録システムと共同するという考えられると思います。

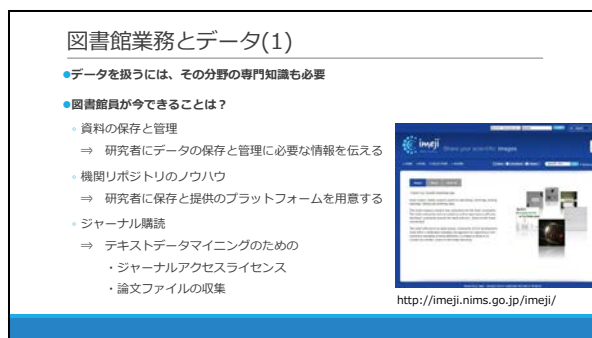
そして、ジャーナル購読のノウハウの利用です。データ駆動型研究では、どれだけデータを集められるかがポイントとなるので、テキスト・データマイニング用に購読とは異なるジャーナルアクセスのライセンスや論文ファイルの収集が必要となります。NIMS で

つい最近、そうした事例があったのでご紹介したいと思います。

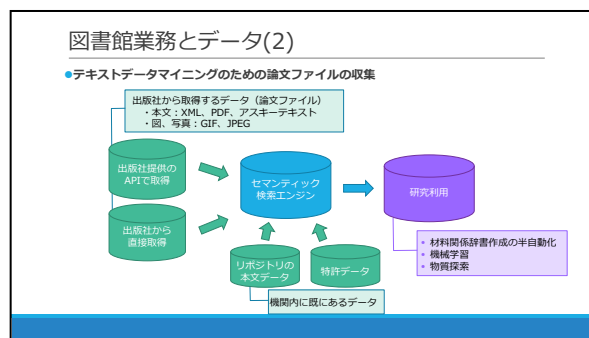
出版社とのコネクションを持っているのが図書館なので、データマイニングのための出版社とのライセンス契約を図書館を通じて行いましたが、出版社によって提供可能な範囲はさまざまでした(図 11)。まず、購読ジャーナルしか認めないというところ、研究目的で NIMS 内だけで使うなら提供できるというところ、研究用途として実験的に無償提供するところ(学会)など、さまざまでした。マイニング用のデータなので、論文 PDF のように美しく成形されている必要はなく、こちらの希望としては本文データと図や写真のデータだけを入手できれば足りるのですが、出版社によって、API 経由だったり、出版社サーバーから直接取得だったり、それぞれ可能な方法でデータを収集しました。

取得したデータを知識ベースとして、セマンティックな検索エンジンに投入しますが、データをよりリッチにするために、機関内で既にあるデータ(リポジトリの本文データや特許データ)についても、このセマンティック検索エンジンに投入します。それを材料関係辞書作成の半自動化や、機械学習、物質探索に利用することを予定しています。

契約交渉を通じてあらためて感じたのは、ライセンス契約の範囲設定の難しさです(図 12)。出版社から論文データを購入する場合、テキスト・データマイニングの成果を機関内にとどめておくのか、それとも第三者に提供するのかで、料金が大きく変わってきます。また、ソフトウェア会社との契約でもセマンティック検索エンジンにより生成した辞書データをソフトウエ



(図 10)



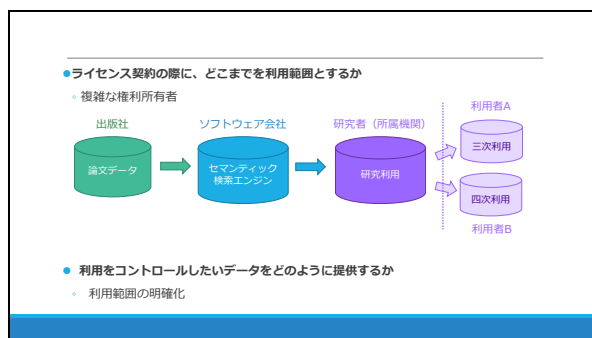
(図 11)

ア会社に提供するかどうかで、やはり料金が変わってきます。ソフトウェア会社に辞書データを提供することは、出版社のデータを第三者に提供することになるので、出版社が第三者への提供を認めていない場合はできません。

NIMS では今回、研究で利用するために NIMS 内だけで限定利用することを前提として、細部については、その都度確認しながら進めるということで、出版社からデータを購入了。NIMS の中だけでの利用、つまり、二次利用までのライセンスですが、研究成果を公開して、三次利用あるいは四次利用ということを見ると、権利の所有者はどんどん複雑になって、ライセンスの内容も、都度、協議するというわけにはいかなることが想定されます。

今回のまとめとして、材料科学分野ではオープンにできるデータとそうでないデータがあり、オープンにできない、つまり、利用をある程度コントロールしたいデータの提供方法の検討ポイントは、利用範囲を明確にすることではないかと思っています。

ライセンスについてはまだまだ勉強中で試行錯誤しているところなので、もし、そういうセミナーを開催される際には、ぜひ教えていただけたらうれしいです。



(図 12)

●角田 JAMSTEC の角田と申します。著作権法をよくよく調べられていて、よく分かりました。ありがとうございます。ただ、そもそも著作権法をいくらじったところで、多分、データそのものをどうこうするのは全く別の話になってしまうのではないかというのが、私がまた別途、弁護士さんに聞いたときの感じでした。

著作権法というのは、中身が事実であろうとシミュレーションであろうと全く関知しないし、中身が面白からうがつまらなからうが、陳腐だろうが独創的だろうが、それも全く関知しません。中身をどう表現するかという表現法が独創的かどうかで保護するというのが著作権法の本来の精神なので、データそのものを著作権法でどうこうしようとしても、それは本質的に無理で、データをどう表現するかに関して、びっくりするような表現方法、独創的な表現方法があるのであれば、それは著作権法の保護の範囲内になるかもしれません。

では、論文はどうかと皆さん思われるのですが、もし論文が本当に定型化されて書かれていて、しかも、どの AI でも、体裁、表現が同じような論文が書けるということであれば、論文の中身がいくら独創的であっても著作権法では保護されないというすごい状況が起きるので、データそのものをどう保護するかということになった場合には、もう著作権法とは全く関係なしにライセンス契約なり何なりを考えるのがいいのではないかと思った次第です。

●小野寺 おっしゃるとおりで、データは著作権で保護される範囲ではないということは、NIMS で開催したセミナーでも弁護士の先生が繰り返しておっしゃっていたことです。データというのは客観的な事実を示したもので、そこに独創性や創造性は関わらないということを学んだところでした。

●**角田** 全くそのとおりなのですが、データにはシミュレーションデータというものもあります。人間が本を書いたり、絵を描いたりするときも、事実をそのまま写すというものと、シミュレーションしたり、頭の中にあることを事実に基づかずに表現するというものもありますよね。そういうこともありますのでということを一言付け加えさせていただきます。

●**小野寺** ありがとうございます。