A chemistry perspective on research data

Richard Kidd
Head of Chemistry Data

27 October 2020

研究データについて
化学からの視点

Richard Kidd
Head of Chemistry Data

27 October 2020

What is research data?

- "*the underlying information researchers measure / process / collect from which they make scientific conclusions*"
- Some fields have clear definitions, e.g. medicine
  - Clinical data: tables of clinical outcomes for individual patients
  - Published paper: statistical summary of clinical trial data + clinical conclusions
  - "*Publish the data associated with this study*" means release the (anonymised) results of individual patients.
  - Obvious separation between "the data" and "the paper".

研究データとは何か?

- **“研究者が科学的な結論を得るため測定・処理・収集する基礎となる情報”**
- 医学など明確な定義のある分野もある
  - 臨床データ:個々の患者の臨床成績の表
  - 公表論文:臨床試験データの統計的要約＋臨床的な結論
  - 「この研究に関連するデータを公表する」とは、個々の患者の(匿名化された)結果を公表することを意味する
  - 「データ」と「論文」は明確に区別される

Research data in chemistry

- “The results” and “the data” are blurred, with limited consensus.
- Synthetic chemistry is particularly challenging
  - Observations, rather than measurements.
  - Limited statistical / numerical analysis
  - Pictures, not words and numbers. No repeat measurements.
  - A “one to many” ratio of techniques to answers.
- Chemists collect “small data” – slowly, manually, and for single systems. Large datasets require aggregation from 1000s of studies
- This is before you think about file formats, standards, identifiers,

化学における研究データ

- 「結果」と「データ」はあいまいで限られたコンセンサスしかない
- 合成化学では特に困難
  - 測定ではなく観察
  - 限られた統計解析/ 数値解析
  - 語句や数字ではなく描写。繰り返しの測定はない
  - 技術と回答が「一対多」の関係
- 化学者はゆっくり、手動で、単一系のために「小さなデータ」を収集する。大規模なデータセットでは何千もの研究を集約する必要がある
- これはファイルのフォーマット、標準、識別子について考える以前のこと

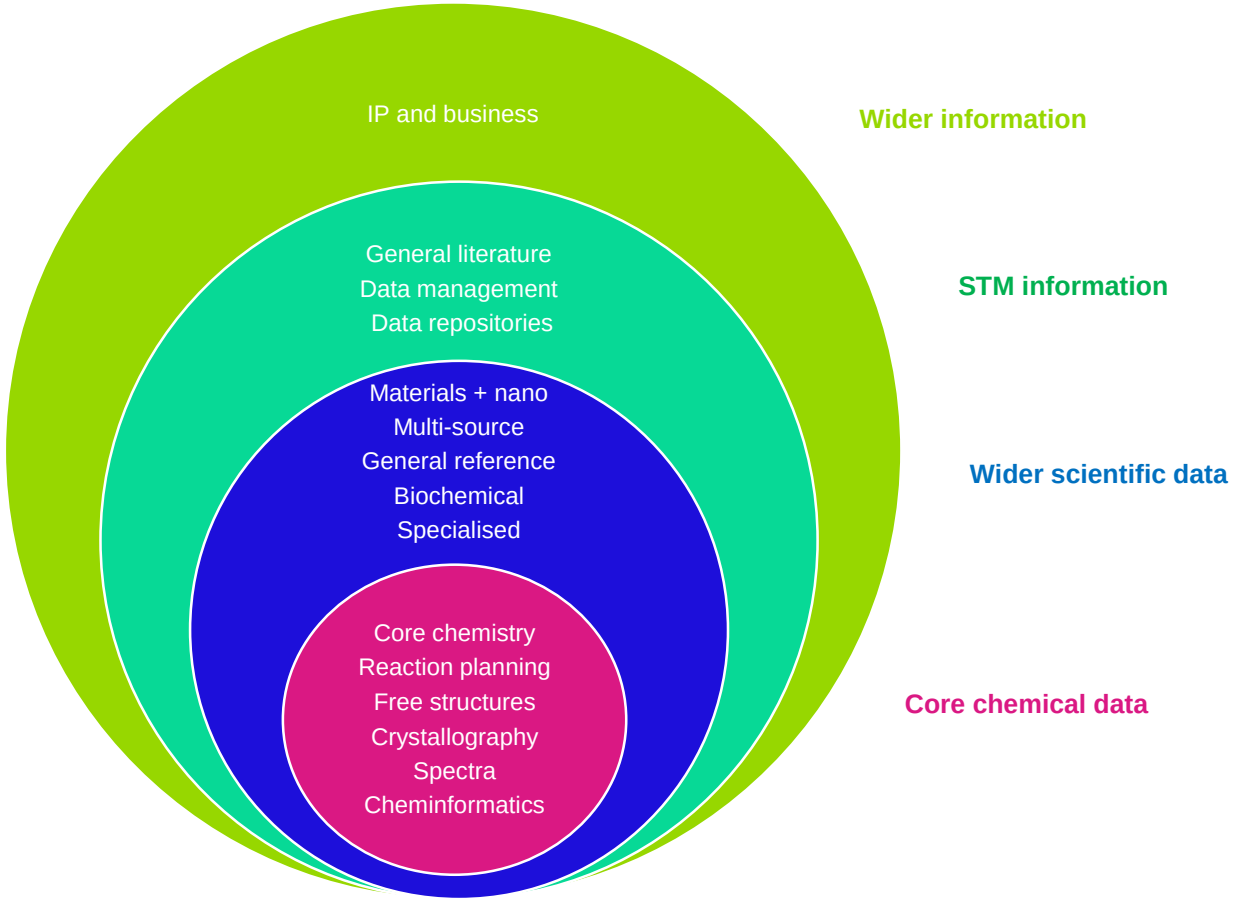
Current publishing practice **in chemistry**

- Supplementary Information (SI) is considered as ‘data’ for most chemists.
- All chemistry publishers mandate crystal data deposition for Cambridge Structural Database (CCDC) and PDB
- All chemistry publishers encourage “full data sharing” while accepting the lack of repositories to do so. Some of them need to improve the clarity and visibility of this message.
- Data linking – in practice means “crystal data linking”
- Data citation - without authors having any data to cite.
- The most practical option is:
- *“The crystal data in this paper are available at [CSD DOI]. All other data is in the SI”*

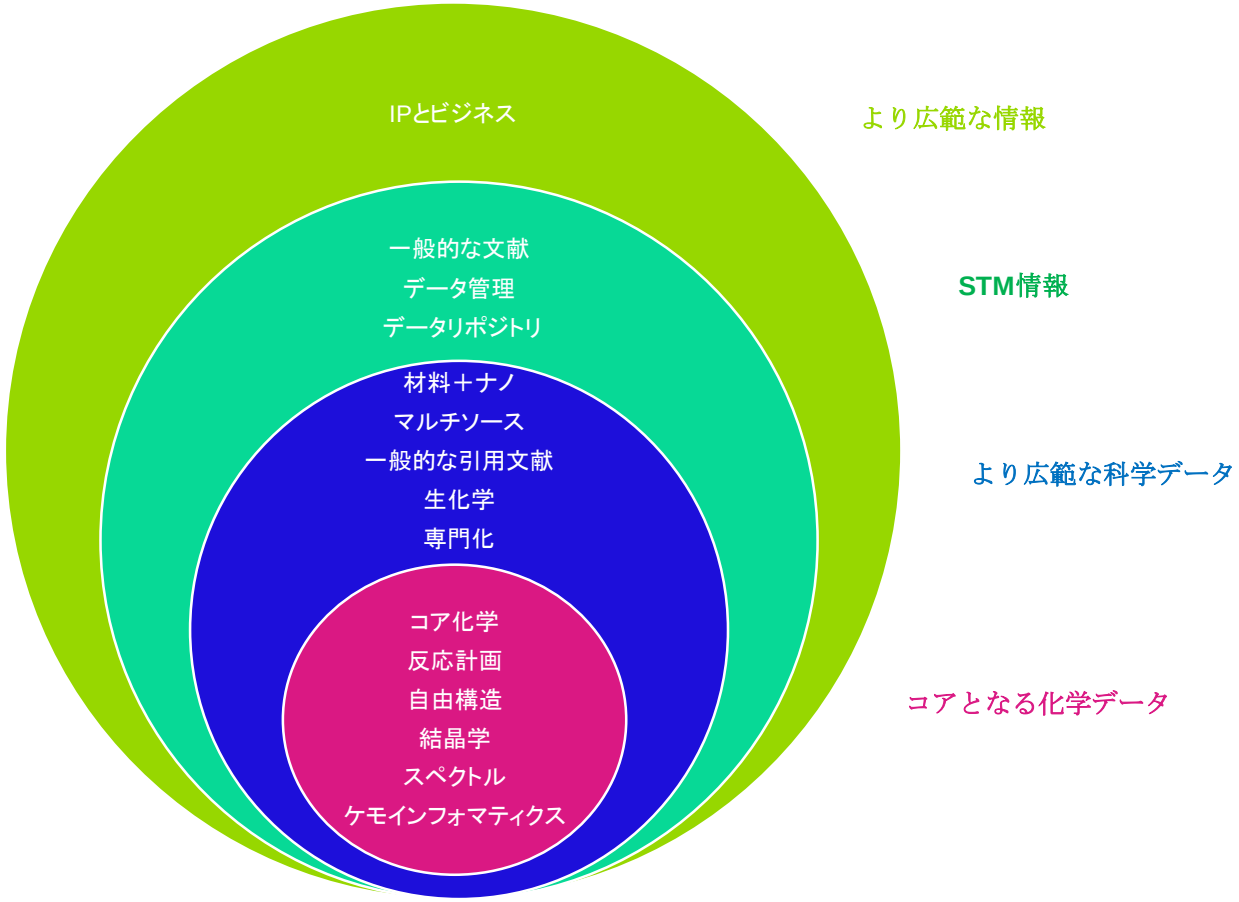
現在の化学における出版活動

- 補遺情報 (SI) は、ほとんどの化学者にとって「データ」と考えられている
- すべての化学出版社は Cambridge Structural Database (CCDC) および PDBへの結晶データの登録を義務づけている
- すべての化学出版社は、そのためのリポジトリの不足を受け入れつつも、「完全なデータ共有」を奨励している。その中には、このメッセージの明確性と可視性を向上させる必要があるものもある
- データのリンクー実際には「結晶データのリンク」を意味する
- データの引用ー著者は引用するデータをもたない
- 最も実用的なオプション:「この論文の結晶データは [CSD DOI] にある。その他のデータはすべてSIにある」

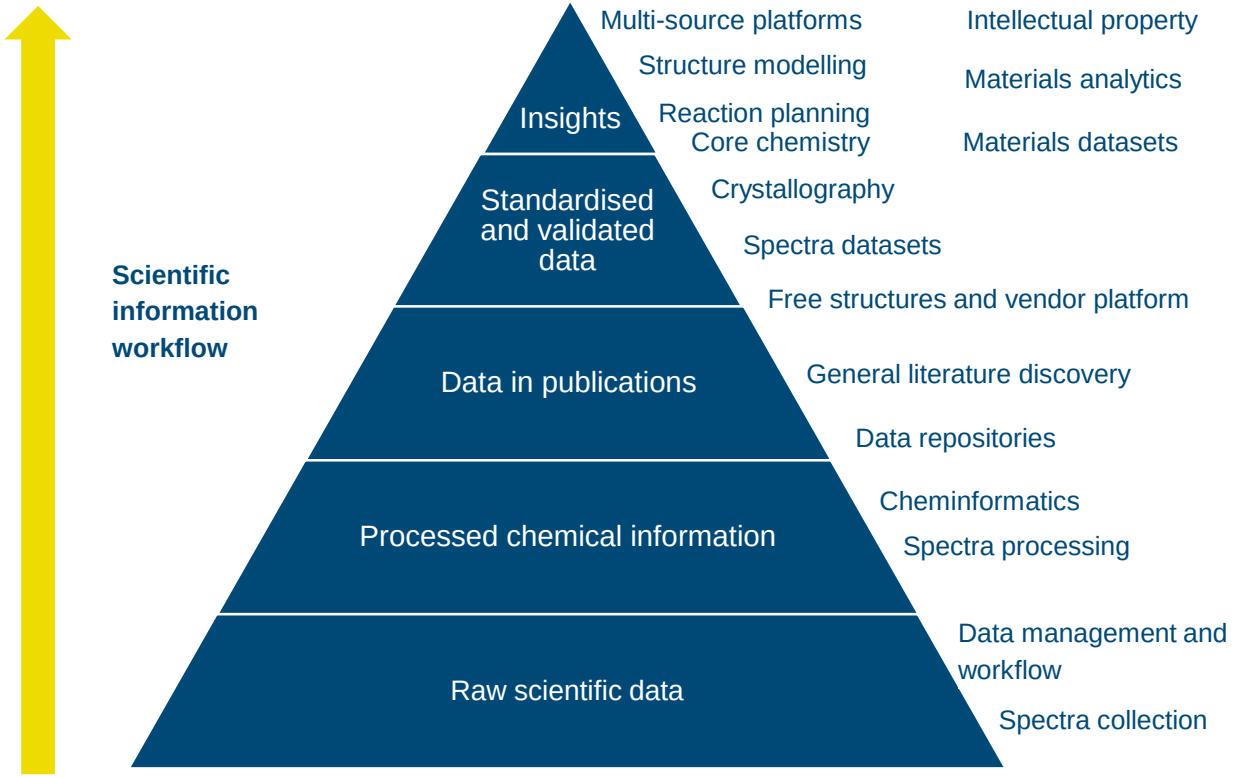
Chemistry data is published (commercially)



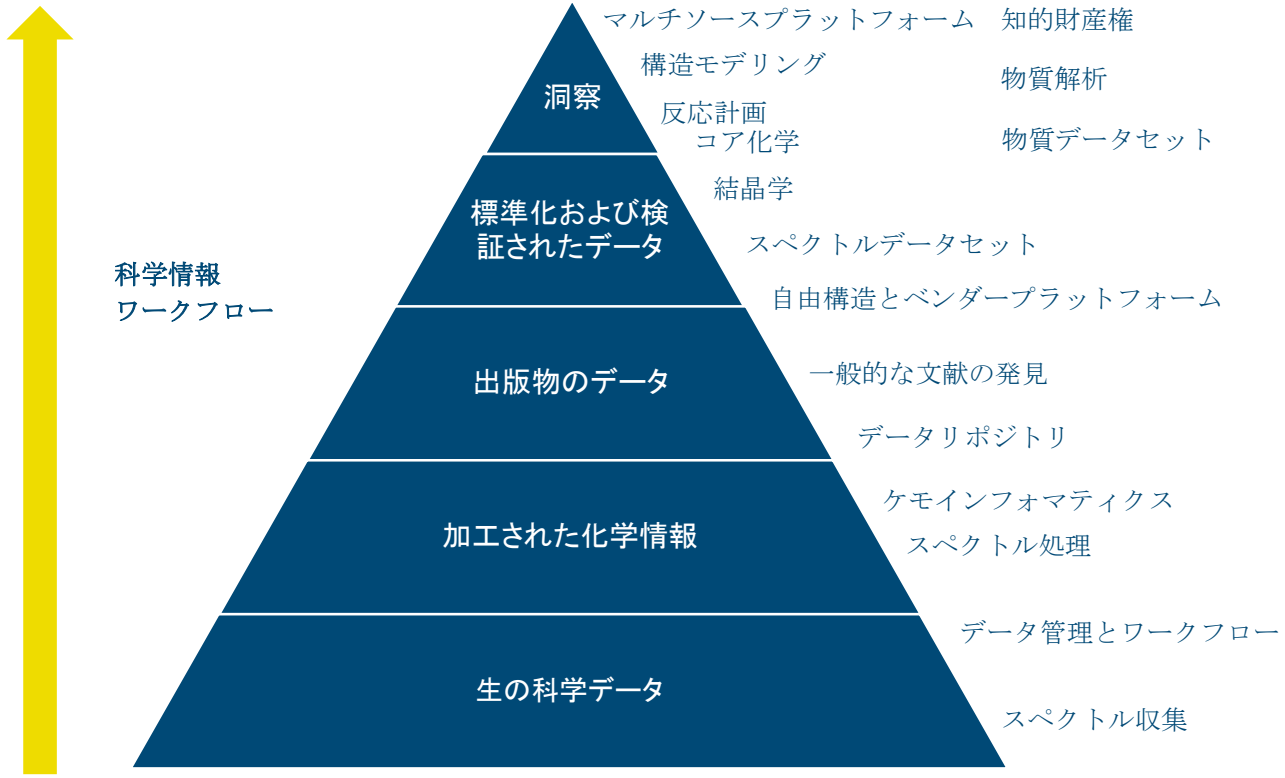
化学データは出版されている（商業的に）

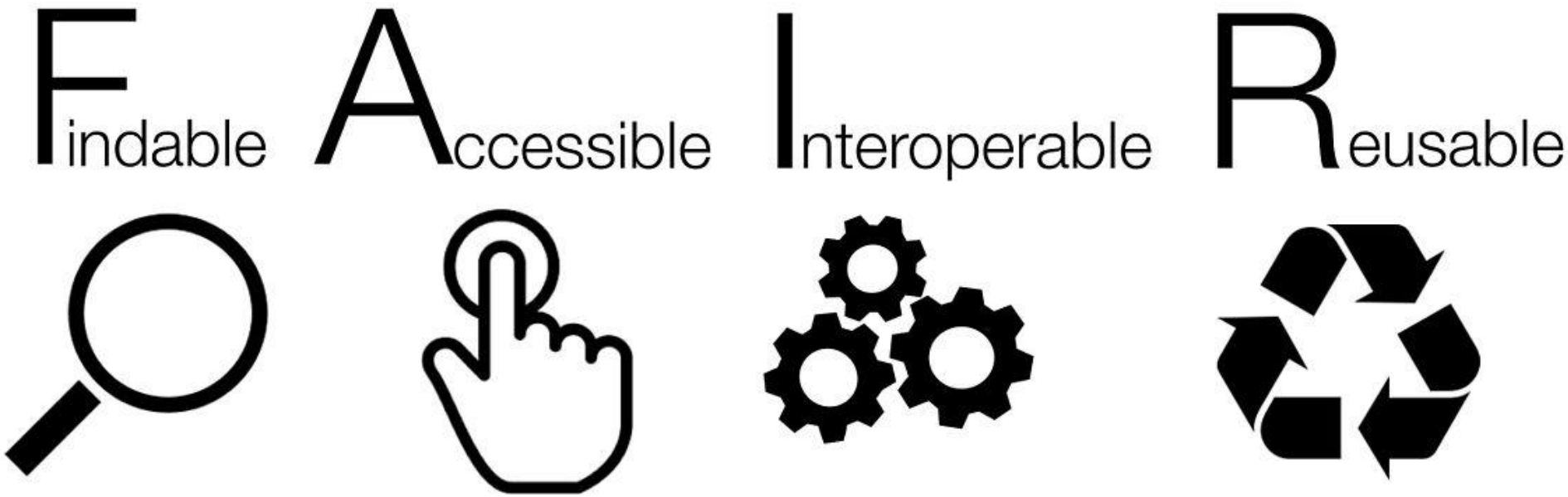


What type of chemical data is this?

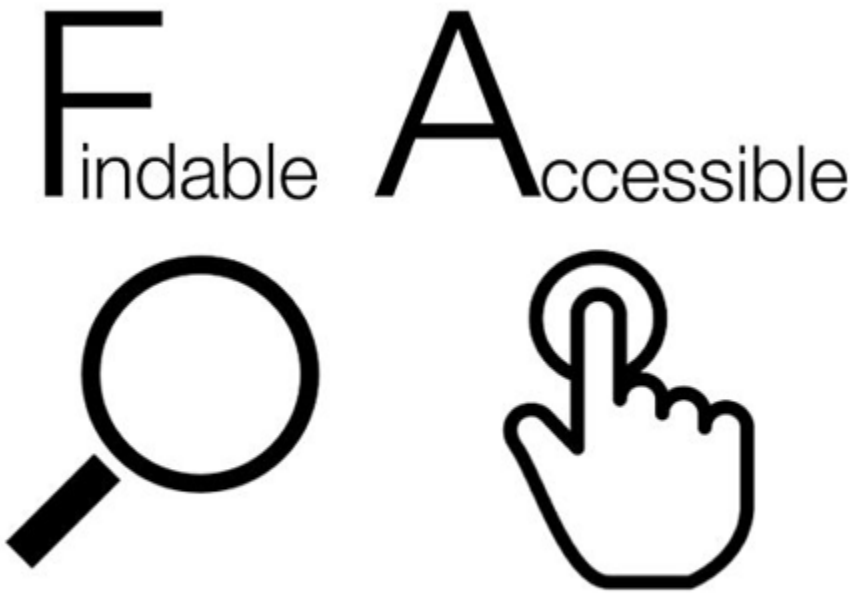


これはどのような種類の化学データなのか?




Findable
Accessible
Interoperable
Reusable

FAIR Challenges



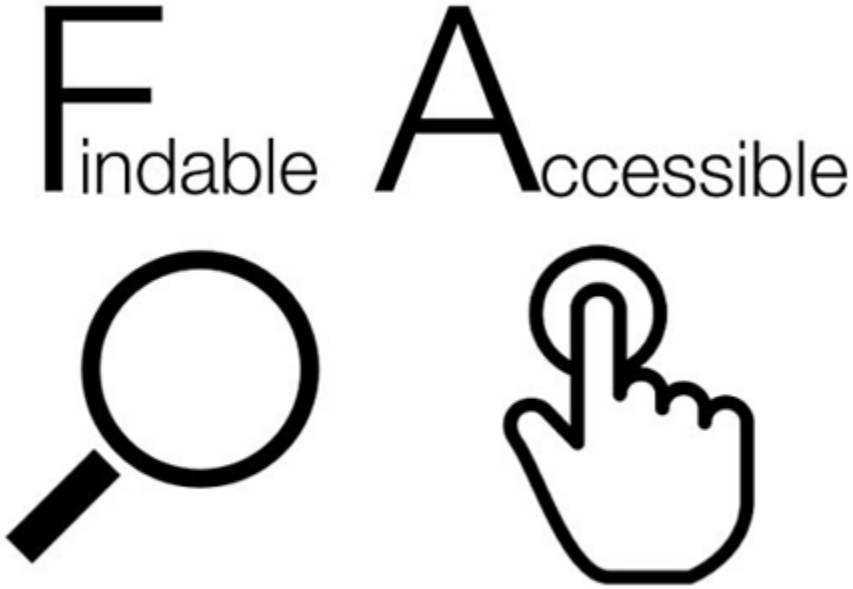
Generally…

SCHOLIX

Data publication standards

Linking publications to datasets

FAIRへの取り組み

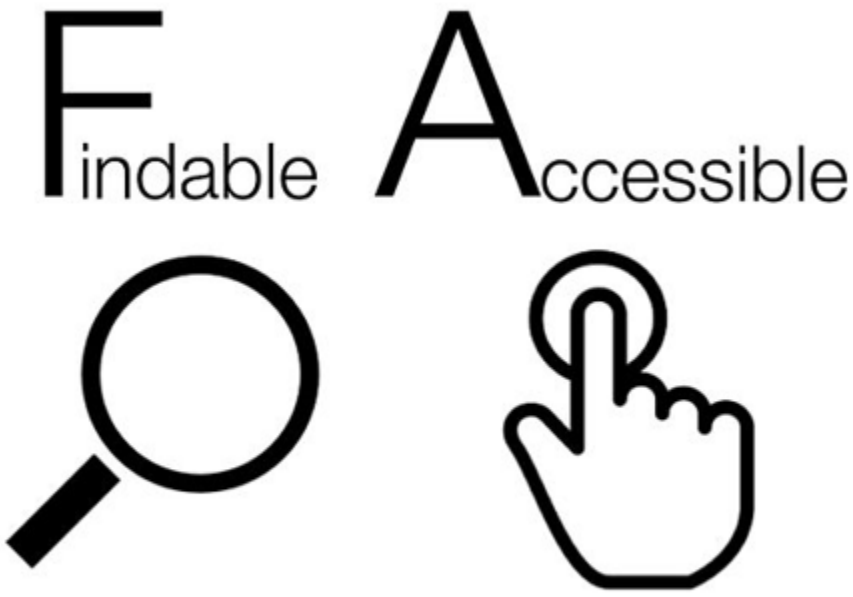


一般的には…

SCHOLIX

データ公開基準
論文とデータセットのリンク

FAIR Challenges for Chemistry



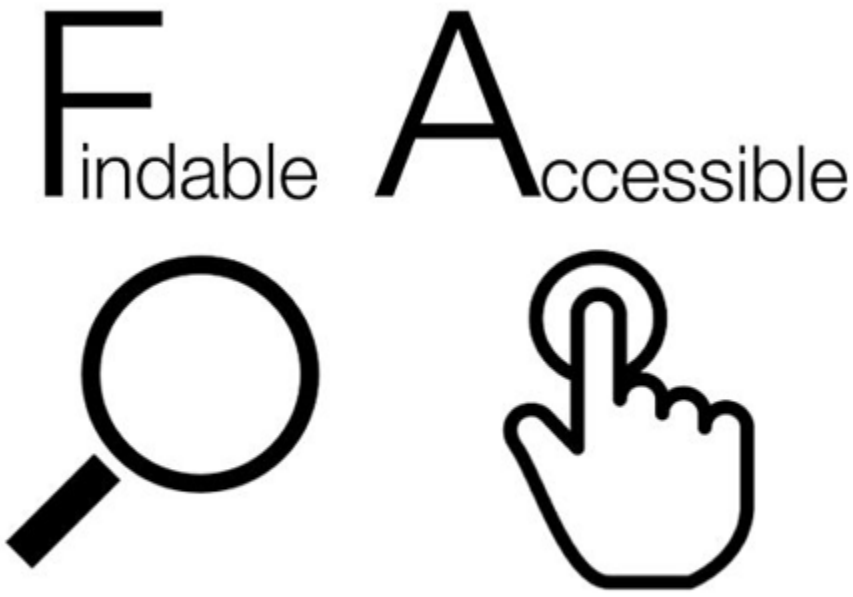
Findability

- Currently no standard representation that handles all of chemistry
- Different approaches to normalisation can have an adverse impact

Accessibility

- Much published chemical data still locked up in text
- Much chemical data proprietary
- Much data captured in value-added closed systems (metadata included)

化学のFAIRへの取り組み



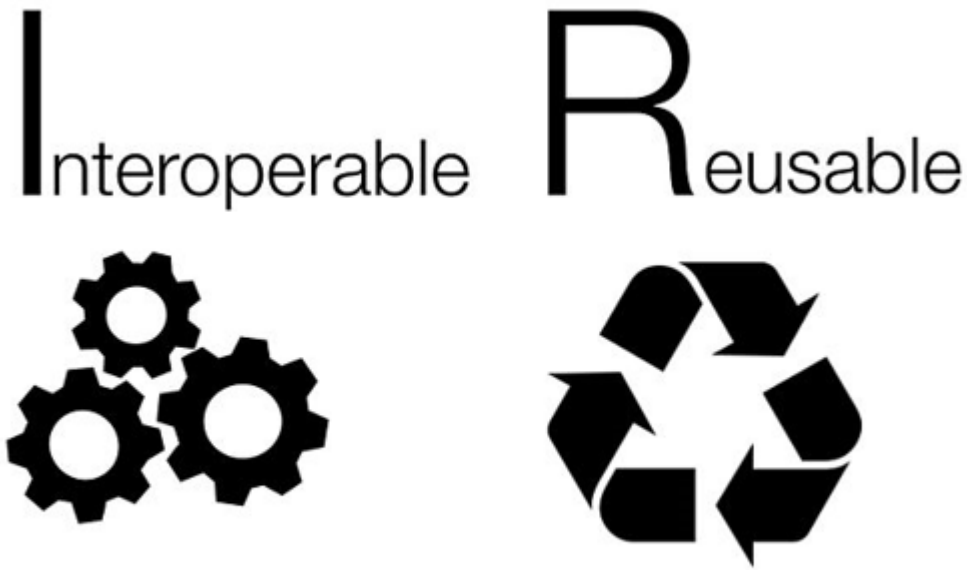
Findability（見つけられる）

- 現在、化学のすべてを扱う標準的な表現はない
- 正規化へのさまざまなアプローチは、悪影響を及ぼす可能性がある

Accessibility（アクセスできる）

- 公開されている多くの化学データはなおテキストにロックされている
- 多くの化学データを専有
- 付加価値の高い閉鎖系に取り込まれた多くのデータ（メタデータを含む）

FAIR Challenges for Chemistry



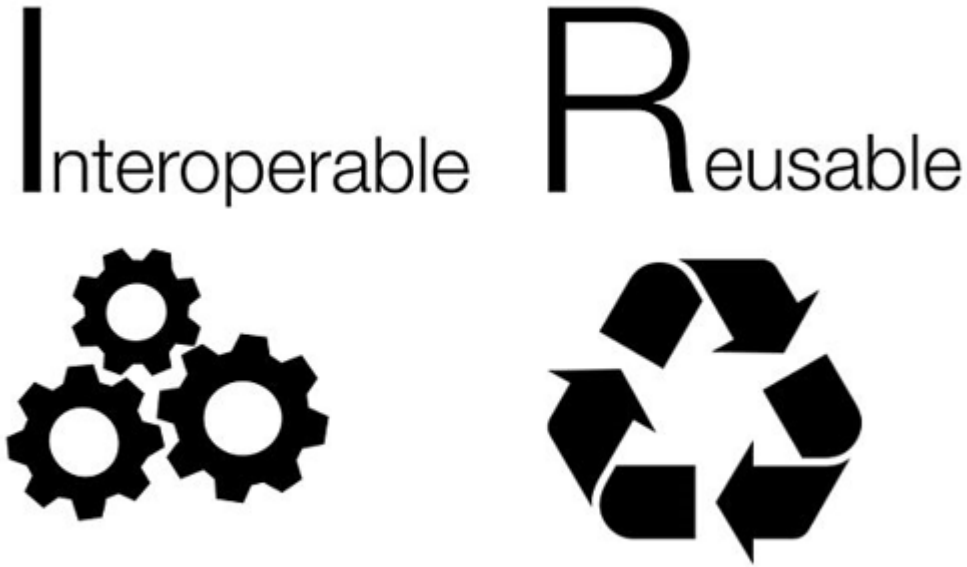
Interoperability

- Tendency for people to locally solve interoperability issues
- Lack of community process to review digital standards
- Insufficient understanding of what other disciplines need from chemistry

Reusability

- There are licenses *but* little guidance/control/choice for the researcher
- Need for training of data producers and consumers to value FAIR for work optimization as well as improved science!

化学のFAIRへの取り組み



Interoperability（相互運用できる）

- 人々が相互運用性の問題を局所で解決する傾向
- デジタル標準をレビューするためのコミュニティプロセスの欠如
- 他の学問分野が化学に何を求めているのかの理解不足

Reusability（再利用できる）

- ライセンスはある。*しかし* 、研究者への指導・管理・選択は少ない
- 科学の進歩だけでなく仕事の最適化にも貢献するFAIRを評価するのに、データ生産者と利用者のトレーニングの必要性

Funder policy in the UK

UK Research & Innovation

“research data should wherever possible be made openly available for use by others in a manner consistent with relevant legal, ethical, disciplinary and regulatory frameworks and norms, and with due regard to the cost involved.”

Concordat on Open Research Data (2016)

英国における資金提供者の政策

英国研究・イノベーション機構

「研究データは、関連する法的、倫理的、学問的、規制上の枠組みおよび規範と整合性のある方法で、また関連するコストを十分に考慮した上で、他の人が使用できるよう可能な限り公開されるべきである」

Open Research Data合意 (2016)

UK Policy - in development

National Data Strategy

Increasing the efficiency and scope of scientific research

"barriers to accessing data represent a significant limitation on research; these range from legal barriers (real and perceived) through to cultural blockers and risk aversion. These barriers must be addressed if the UK is to remain at the forefront of science and research. For example, recent research into data use by the pharmaceutical and life sciences industry identified a number of systemic barriers that limit access to data. Most companies surveyed noted having experienced delays and uncertainties. These include time taken to access data, access constraints for commercial users, the effort to identify and assess the quality of data sets and, most notably, the cost of the data itself."

https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy

英国の政策ー策定中

National Data Strategy

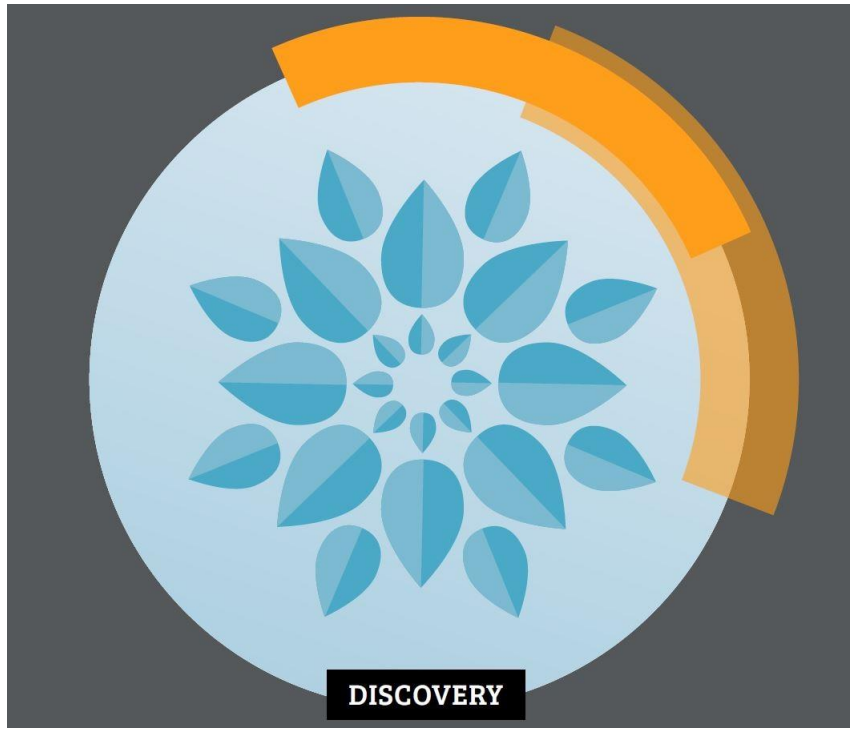
科学研究の効率と範囲を拡大する

“データのアクセスへの障壁は研究の大きな制限となっていて、これらの障壁は、（実在する、および、感知される）法的な障壁から文化的な障壁、リスク回避に至るまで多岐にわたる。英国が科学と研究の最前線にとどまるには、これらの障壁に対処しなければならない。たとえば、製薬・ライフサイエンス業界によるデータ利用に関する最近の調査では、データへのアクセスを制限する多くの体系的な障壁が確認された。調査対象となったほとんどの企業は、遅延や不確実性を経験したと回答している。これらには、データへのアクセスにかかる時間、商用ユーザーのアクセスに対する制約、データセットの品質を判断し評価するために必要な労力、そして何よりもデータ自体のコストが含まれる”

https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy

Science Horizons

Prospects for advances in academic research in next 10-15 years

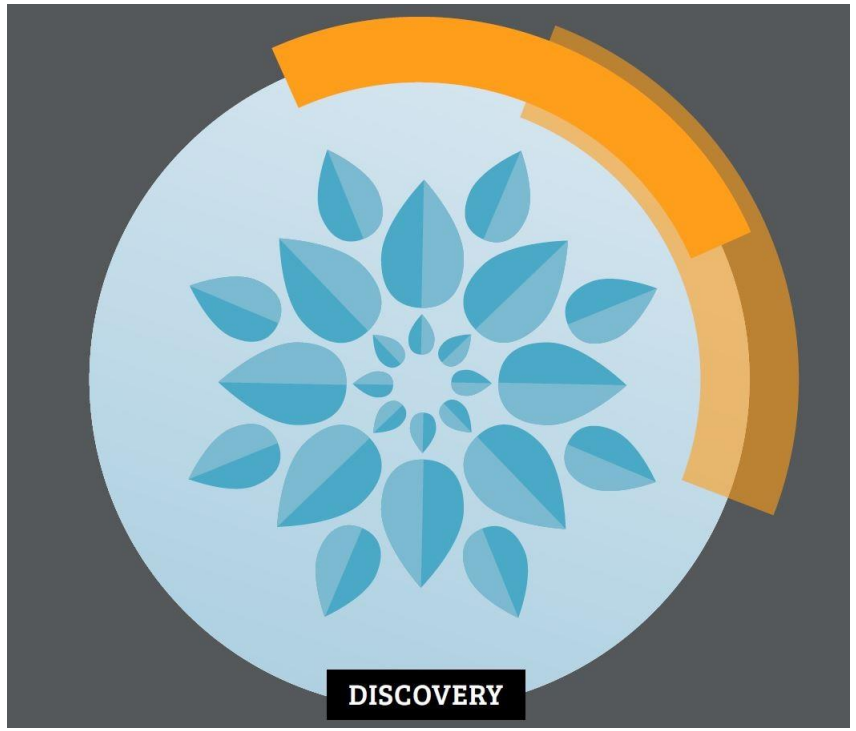


ROYAL SOCIETY OF CHEMISTRY

rsc.li/sciencehorizons
(published summer 2019)

Science Horizons

今後10～15年の
学術研究の進展の見通し



ROYAL SOCIETY OF CHEMISTRY

rsc.li/sciencehorizons
(published summer 2019)

Science Horizons: Summary Outlook

1. Global **R&D investment strong** [check post-covid]
2. **Sustainability & prosperity** take centre stage in R&I policy

3. Researchers are committed to delivering **solutions to global & industrial challenges**
4. Researchers continue the scientific quest to **answer fundamental questions**
5. **New techniques** are opening new frontiers in scientific discovery & application
6. **Data and digital technologies are integral to research and create exciting new possibilities**
7. Interplay between **theory, experiment & computation** is crucial, especially in new areas

8. Research is **international**
9. Research is increasingly **interdisciplinary** and this needs support from funders & universities
10. **Curiosity and fundamental discovery science** are important – in their own right and for innovation – and need more sustained long-term funding
11. **Collaboration with industry** is key, and we need ways to make this straightforward for everyone
12. **Leadership** within and on behalf of science are crucial

ROYAL SOCIETY OF CHEMISTRY

Science Horizons: 概観

1. 世界規模で**研究開発投資が強化**（ポストコロナ）
2. **持続可能性と繁栄**がR&I政策の中心的な役割を果たす

3. 研究者は、**世界全体および産業界の課題に対するソリューション**の提供に尽力している
4. 研究者は**根本的な疑問に答える**ために科学的探求を続ける
5. **新しい技術**が科学的発見と応用の新境地を切り開く
6. **データとデジタル技術は研究に不可欠であり、刺激的な新しい可能性を生み出す**
7. 特に新しい分野では、**理論、実験、計算**の相互作用が非常に重要である

8. 研究は**国際的**なものである
9. 研究はますます**学際的**になってきており、資金提供者や大学からの支援を必要とする
10. **好奇心と基礎的発見の科学**は、それ自体とイノベーションのために重要であり、より持続的な長期の資金調達が必要である
11. **産業界との連携**は重要であり、これを誰にとってもわかりやすくする方法が必要となる
12. 科学の中で、そして科学を代表して**リーダーシップ**を発揮することは非常に重要である

ROYAL SOCIETY OF CHEMISTRY

Data and digital technologies in Science R&D – why does this matter?

1. An unanswered question: How transformative is the use of data and digital technologies in science? Is this a revolution or is it hype?

2. A new perspective on a hot topic: Not just the ‘usual’ applications of AI & robotics like manufacturing, recommending products, image recognition.

3. Bringing something hidden to the surface: Not just AI & robotics, and not all new to scientists. Also modelling & simulation, measurement & sensing, imaging & visualisation

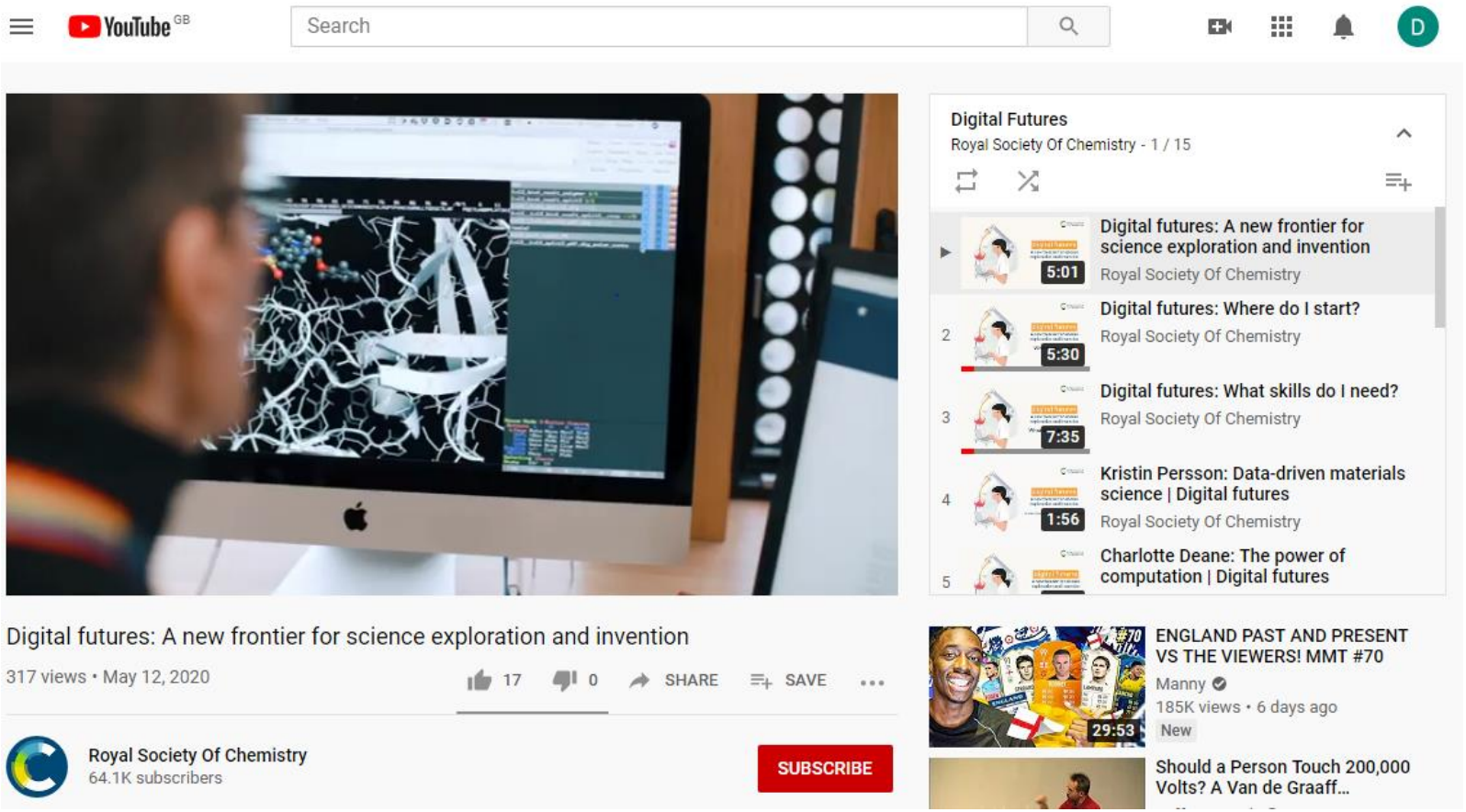
科学研究開発におけるデータとデジタル技術 – なぜそれが重要なのか？

1. 正解の無い問い：科学におけるデータとデジタル技術の利用はどのように変容しているのか？これは革命なのか、それとも誇大広告なのか？

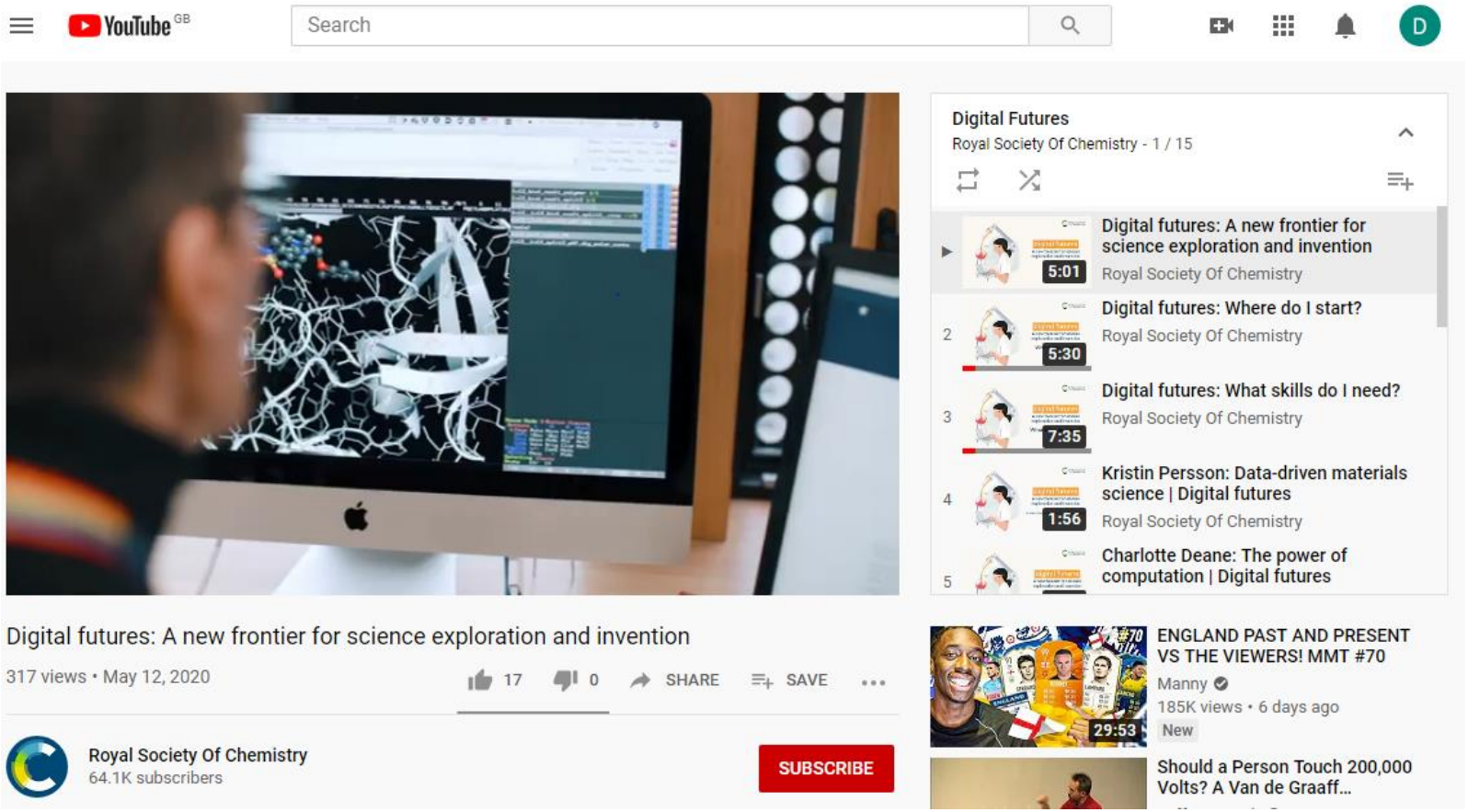
2. ホットな話題における新しい視点：製造業、製品の推薦、画像認識のような、「普通の」AI＆ロボティクスのアプリケーションだけではない。

3. 隠れた何かを表面化させる：AIやロボティクスだけではないし、科学者にとって目新しいものばかりではない。モデリングとシミュレーション、計測とセンシング、イメージングと可視化、など

Digital Futures launch (July 2020)



Digital Futuresの開始 (2020年7月)



Will data & digital tech transform Science R&D?

What will be really disruptive is how digital tools will push us to develop new or different ideas and hypotheses, they will amplify human creativity, creating new possibilities or opportunities

We finally have the beginnings of a revolution in materials science. We can use quantum mechanics and first-principles simulations to predict novel materials, but we need robotics and automation and to use a closed-loop system. Otherwise we won't fail fast enough to find new battery cathode materials for the future.

People wanted to probe the entire cell before, but couldn't do it. Now we have the tools and techniques to answer the bigger question. Eventually we might probe the entire human – not in the next 10-20 years, but it will happen one day.

Computers now can hold far more information and complexity, and explore more possibilities, than an individual human brain. We need to harness this to help us see patterns and options that a human being would not find

データとデジタル技術は科学研究開発を変えるか？

本当に破壊的なのは、デジタルツールがいかに私たちに新しいアイデアや仮説の開発を促し、人間の創造性を増幅させ、新しい可能性や機会を生み出すかということです。

私たちはついに材料科学の革命の始まりを手に入れたのです。量子力学と第一原理シミュレーションを使って新しい材料を予測することができますが、ロボット工学と自動化、そして閉ループ系を使う必要があります。そうでなければ、将来の新しい電池正極材料を見つけるのに十分な速さで失敗することはないでしょう。

以前は細胞全体を調べようとしていましたが不可能でした。今、私たちは大きな疑問に答えるためのツールと技術を持っています。最終的には人類全体を調査することになるかもしれません - 今後10年から20年とはいきませんが、いつかは実現するでしょう。

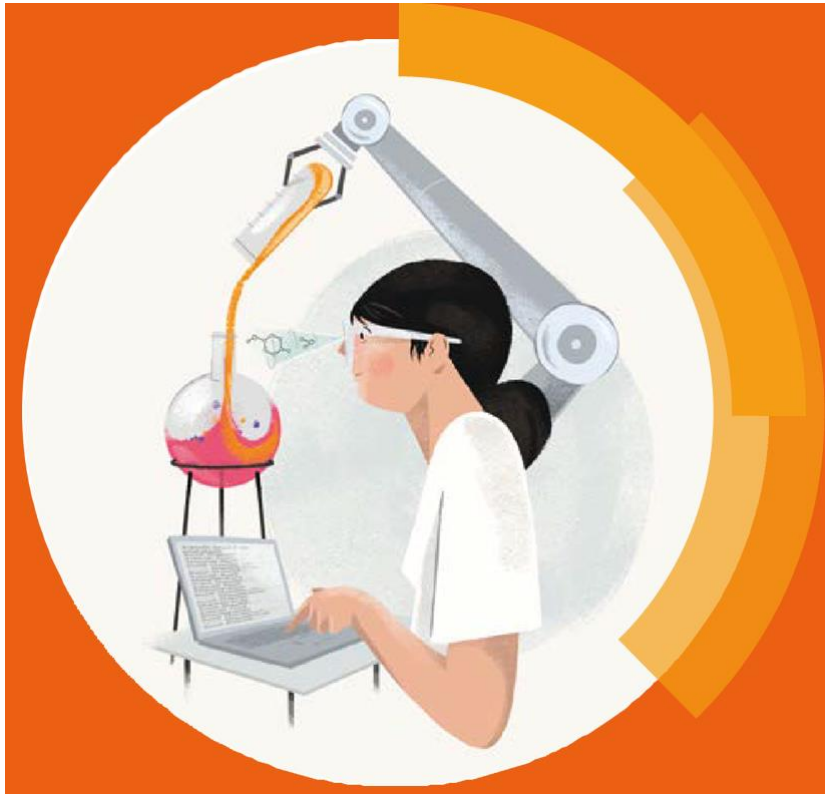
- *今やコンピュータは、人間の脳よりもはるかに多くの情報と複雑さを持ち、より多くの可能性を探求することができます。私たちは、人間が見つけられないようなパターンや選択肢を見る補助手段としてこれを利用する必要があります。*

Data within a Digital Future

Core to chemistry is the synthesis and characterisation of substances – and how these substances interact from the sub-molecular to the global level. Digital technologies [..] will increase efficiency and sustainability across the chain from sourcing raw materials, to product development and manufacturing, to distribution, consumption and end of product life.

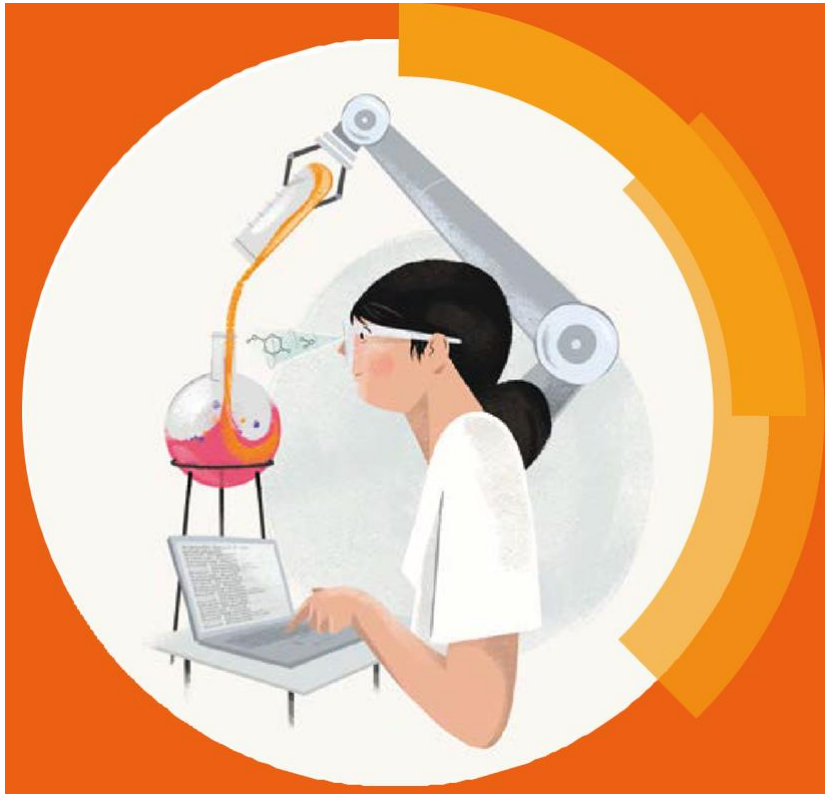
Data is a key area of technical opportunity – bringing together and sharing data from different sources to enable better, faster, bigger science

To support this future will need incentives for data sharing; development of common standards and formats; and the availability of infrastructure, tools and skills for the curation and use of high-quality datasets.

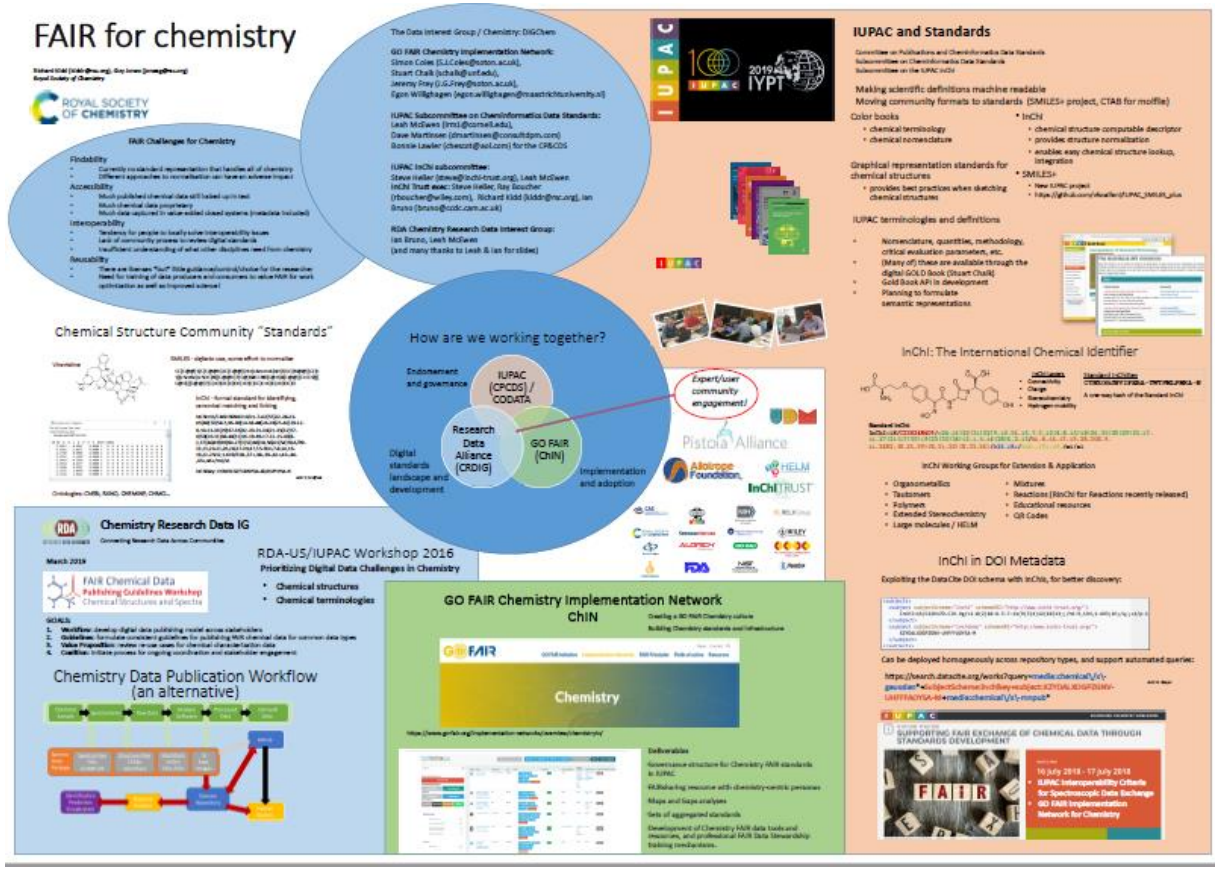


Digital Futureにおけるデータ

- 化学の中核となるのは、物質の合成と特性評価であり、これらの物質が分子レベルからグローバルレベルまでどのように相互作用しているかを知ることです。デジタル技術は、原材料の調達から製品開発、製造、流通、消費、製品寿命の最終段階に至るまでのチェーン全体の効率性と持続可能性を向上させます。
- データは技術的な機会の重要な分野です。さまざまなソースからデータを集めて共有することで、より良い、より速い、より大きな科学を可能にします。
- このような未来をサポートするために、データ共有のためのインセンティブ、共通の標準規格とフォーマットの開発、高品質なデータセットのキュレーションと利用のためのインフラストラクチャ、ツール、スキルの利用可能性が必要となります。



Different examples for RSC



ChemSpider
Search and share chemistry

ChemRxiv™

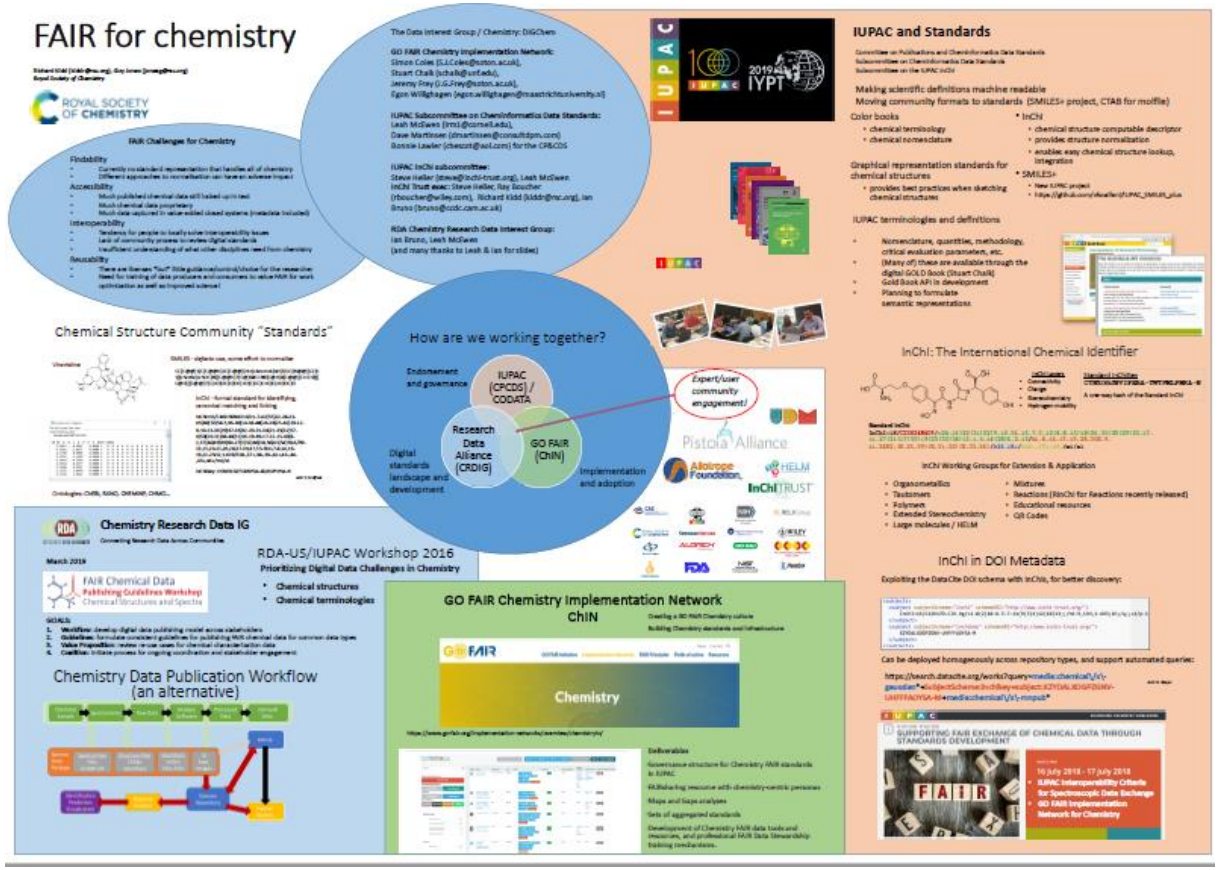
ACS Chemistry for Life®

CCS 中国化学会 CHINESE CHEMICAL SOCIETY

THE CHEMICAL SOCIETY OF JAPAN

GDCh GESELLSCHAFT DEUTSCHER CHEMIKER

RSCの様々な例



ChemSpider
Search and share chemistry

ChemRxiv™

ACS Chemistry for Life®

CCS 中国化学会 CHINESE CHEMICAL SOCIETY

THE CHEMICAL SOCIETY OF JAPAN

GDCh GESELLSCHAFT DEUTSCHER CHEMIKER

Data generation and management

- What is “the data” and how to collect/aggregate/process it?
- Is the data on file, collected and processed as part of authors’ current work (writing papers). Is it close to being ready for others to use?

- Standards come with a mature field

データの生成と管理

- 「データ」とは何か、それをどのように収集/集計/加工するのか？
- 著者の現在の研究（論文執筆）の一環として収集・処理されたデータはファイルに保存されているか、他の人が利用できる状態に近いか？

- 成熟した分野であれば標準化されている

Where to put data?

General repositories

Domain repositories

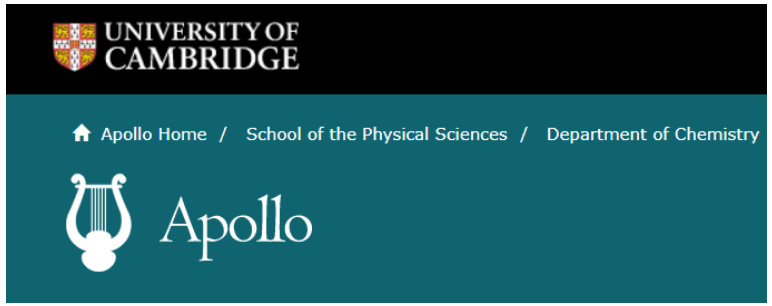
CSD 2020.2

データをどこに置くか？

一般的なリポジトリ

figshare

zenodo



ドメインリポジトリ

CSD 2020.2

RCSB PDB PROTEIN DATA BANK

ChemSpider
Search and share chemistry

PubChem

The RSC's data policy

- We’re considering data availability statements
- We’ll add data citations
- Author incentives (and data availability, formats and domain repositories) are lacking – what can we do to support?
- Look for subject opportunities

RSCにおけるデータポリシー

- Data availability statements（データの利用可能性に関する記述）を検討している
- データ引用を追加する予定
- 著者のインセンティブ（およびデータの利用可能性、フォーマット、ドメインリポジトリ）が不足している - 何を支援することができるか？
- 課題の機会を探す

How to prepare for open data

- What is important to your community
- Findable and Accessible are relatively solved (if the authors have the data)
- Interoperable and Reusable is dependent on the domain, and will evolve.

オープンデータを準備する方法

- 当該コミュニティにおいて何が重要なのか
- （著者がデータを持っている場合）“見つけられる”と“アクセスできる”は比較的解決されている
- “相互運用できる”と“再利用できる”はドメインに依存しており、今後も進化していくだろう

Future definition of chemistry data ?

Machine-readable chemical information

- Structured information / standard formats
- Data with context / metadata
- Bias towards FAIR
  (Findable, Accessible, Interoperable, Reusable)

化学データの将来的な定義?

機械可読な
化学情報

- 構造化された情報/標準的なフォーマット
- 背景情報/メタデータを備えたデータ
- FAIR(見つけられる, アクセスできる, 相互運用できる, 再利用できる)原則の偏重

Thank you

- kiddr@rsc.org

ありがとうございました

- kiddr@rsc.org