eScience Infrastructures in Physical Chemistry

Samantha Kanza[1,2] Cerys Willoughby[1,3] Colin Leonard Bird[1,4] and Jeremy Graham Frey[1,5]

[1]School of Chemistry, University of Southampton, SO17 1BJ, UK
[2]s.kanza@soton.ac.uk, https://orcid.org/0000-0002-4831-9489
[3]cerys.willoughby@soton.ac.uk, https://orcid.org/0000-0003-1721-9212
[4]colinl.bird@soton.ac.uk, http://orcid.org/0000-0003-2154-8158
[5]j.g.frey@soton.ac.uk, https://orcid.org/0000-0003-0842-4302

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–22

https://doi.org/10.1146/((please add article doi))

Keywords

eScience, data, semantic web, collaboration, open science, digital

Abstract

As the volume of data associated with scientific research has exploded over recent years, the use of digital infrastructures to support this research and the data underpinning it has increased significantly. Physical chemists have been making use of eScience infrastructures since their conception, but in the last five years their usage has been even greater. Whilst these infrastructures have not greatly affected the chemistry itself, they have in some cases had a significant impact on how the research is undertaken. The combination of the human effort of collaboration to create open source software tools and semantic resources, the increased availability of hardware for the laboratories, and the range of data management tools available has made the life of a physical chemist significantly easier. This review considers the different aspects of eScience infrastructures and explores how they have improved the way in which we can conduct physical chemistry research.

Contents

1. INTRODUCTION

We are living in an increasingly digital era, where technology underpins almost every aspect of our lives. It is unsurprising therefore that many jobs and research areas, including the physical sciences, now heavily utilize, and in many cases, completely rely on digital infrastructures. A key set of infrastructures that have made an immeasurable difference to the ease and capability to conduct research in the physical sciences, are those of eScience (1).

This review looks to explore how the technologies, methodologies, and collaborations enabled by eScience have improved the way physical chemistry is conducted in the third decade of the 21^{st} Century. A brief history of eScience will be given, followed by a description of the current state of physical chemistry. We then describe the three main themes that have emerged in the context of eScience infrastructures for physical chemistry: collaboration, data & data management; and the use of novel technical methodologies. The article concludes with the authors' perspective on how eScience infrastructures have shaped and enhanced how physical chemistry is conducted in the 21^{st} Century.

2. THE ESCIENCE STORY: SETTING THE SCENE

In 2020, eScience reached its 21^{st} birthday, albeit after an indeterminate gestation period. The term was introduced in the UK by John Taylor in 1999, and was used to title the large UK programme launched in 2000, as related in the comprehensive Wikipedia article (2). There are numerous descriptions of eScience or e-Science, of which the IGI-Global definition (3) captures the components that contribute to effective eScience Infrastructures.

> *"... scientific research based on the collaboration within a number of scientific areas, enabled by a next generation infrastructure, wherein people, computing resources, data and instruments are brought together to bring a new quality to the everyday work of researchers.*

Related programmes, mainly in the US, were varyingly called cyber-science, or cyber-infrastructure initiatives. The two features of eScience Infrastructures that have substantially facilitated 21^{st} century advances in chemistry and other sciences are data management and networks. John Wilbanks sums up the importance of data management and networks of people aptly in his article in The Fourth Paradigm (4):

But there is precious little in terms of alternatives to the network approach. The data deluge is real, and it's not slowing down. We can measure more, faster, than ever before. We can do so in massively parallel fashion. And our brain capacity is pretty well frozen at one brain per person. We have to work together if we're going to keep up, and networks are the best collaborative tool we've ever built as a culture. And that means we need to make our data approach just as open as the protocols that connect computers and documents. It's the only way we can get the level of scale that we need.

The importance of collaboration, while long recognised, was to become a prominent feature of eScience history. Early initiatives, such as Publication at Source (5) sought to enable expeditious take-up of research findings by other workers, thus making it easier to collaborate, record, and to carry out science in general.

As laboratories have become more automated, with data generated electronically via laboratory instruments, the volume of data produced by scientific research has been increasing exponentially over the years. Thus, there has been a need for both the appropriate computing resources and the improved data management processes to handle this. One of the early infrastructures employed by eScience to deal with substantial volumes of data and large-scale collaboration was grid computing (6), which was employed by many early eScience projects (7, 8, 9). However, as we have moved into the 21^{st} Century, clouds and cloud computing have become dominant. These two services offer similar affordances but use different mechanisms. Khillar has written an article describing the difference between the two concepts (10), in which he offers the following distinction: "In grid computing, resources are distributed over grids, whereas in cloud computing, resources are managed centrally." However, ultimately the main purpose of these resources was to enable users to store and process large volumes of data.

It became apparent that merely being able to store these large volumes of data was not enough. In order to maximise its use, there was a clear need to develop mechanisms for exchanging and managing the vast amounts of data and information. Early eScience projects such as CombeChem (11, 12) looked at using Semantic Web technologies to describe and link together diverse datasets in chemistry. If we fast forward to today, researchers are still making use of these technologies for data representation, and are beginning to use "sophisticated machine learning and other AI technologies both to automate parts of the data pipeline and also to find new scientific discoveries in the deluge of experimental data." (13). However, there is still plenty of work to be done to facilitate capturing the provenance and enabling discovery of physical chemical data.

There are many more aspects to eScience. However, surveying its evolution is outside the main remit of this article, whose main purpose is to describe how eScience has improved the way we are able to do physical chemistry, and to identify the main areas in which these improvements have been made. Note that we will not attempt to cover the ways in which increased computational power and the consequent increase in the range of applications of computational chemistry have impacted on physical chemistry. We will also not cover in detail the advances in Artificial Intelligence (AI) and more specifically Machine Learning (ML) in physical chemistry, despite the major changes these may bring. Nevertheless, we note that AI and ML succeed only when presented with large amounts of quality information and as will become clear the eScience technologies have led to major improvements in the quantity, quality and accessibility of such chemical data.

3. PHYSICAL CHEMISTRY & ESCIENCE IN THE THIRD DECADE

Physical Chemistry covers a very broad range of topics and approaches, spanning theoretical and computational work by researchers who never enter a laboratory, through laboratory bench scale experiments, to large laboratories and experiments conducted in or via large-scale facilities. A characteristic of much of physical chemistry is data analysis – the lab-based experimentalists will typically spend as much or more time in the analysis of their data as collecting it (unless they are involved in building new instrumentation). Even those running large-scale simulations will be involved in significant data analysis of the simulation results.

Historically data has been collected manually and on a relatively small scale, but with increasingly large amounts of data being captured in digital form, there are challenges to the management and analysis of such data. Large-scale experiments may produce so-called 'Big Data', where the sheer volume and complexity of the data causes challenges in storing, transporting and extracting the data to analyse. Even small-scale data presents challenges when attempting to compare, replicate, and reuse data from the wider scientific community. It is well known that researchers are often reluctant to share their data with others, making the vast majority of data inaccessible (14).

Over time, data can become lost or incompatible with the latest software. Even when data is made available it is often incomplete, in a format that is unsuitable for reuse or comparison with other data, and missing important contextual information, making it impossible to understand or use. Data is often separated from the processes and other tools that were used to generate those results. Data is often collected without a specific plan as to how it will be stored and managed at the end of the project, and little consideration is given to how it could be of use to others outside of the research group. However, building upon the work of others and implementing the advanced computing techniques to solve scientific grand challenges requires both access to and utilisation of such data.

Therefore, in considering the impact of eScience infrastructures on Physical Chemistry we need to consider several interrelated aspects:

- **Collaboration & The Social Machine of Physical Chemistry:** The Web, and therefore eScience infrastructures, which are frequently built on web-based technologies, is more than just a technology, it is a socio-technical phenomenon (15). eScience is a Social Machine that comprises data, computers, and people. It is the people who have been driving both the development of the infrastructures and the adoption of new infrastructures that help to enhance how we do physical chemistry.
- **Smart Laboratories & Remote Experimentation:** There has been a rise of the use of Internet of Things (IoT) sensors, pervasive computing and smart instruments in the laboratory that make conducting experiments, collecting data and monitoring different aspects of the laboratory easier. Remote experimentation at facilities (e.g. synchrotrons, high powered lasers, neutron facilities, telescopes), has a moderately long history. Only a few members of a research group are required to be at a facility with others staying 'at home' and accessing the facility remotely. Im some cases the whole experiment may be conducted completely remotely, as is the case for some astronomy research. Additionally, COVID-19 has brought considerations of remote lab work much closer to most chemists. The need to reduce the number of people in a laboratory and still provide adequate supervision, instruction and advice has led many more bench scientists to investigate remote connections, video links, and other aspects of computer aided research that are part of the eScience culture.

- **Data Collection, Storage & Curation:** There are key aspects of data management that have improved how we do physical chemistry, ranging from data collection and generation techniques, to how we store the data and in what form, and how we curate data to enhance its availability to other scientists. In a paper about curation in the chemical sciences, we have argued that curation is most effective when carried out at source, when the data is collected (16).
- **Data Access & Sharing:** Data management also entails consideration of how data will be used and shared with other researchers, to which end platforms have been created to facilitate the conduct of physical chemistry research. Some of these are domain specific, and others are more general tools to store data and experimental information (Electronic Lab Notebooks), to analyse data (Jupyter Notebooks (17)) and to publish results (LaTeX & Overleaf (18)).
- **Knowledge Representation & Elicitation:** The use of Semantic Web Technologies in Physical Chemistry research has vastly increased over the last few years, with researchers developing new ways to represent and model the data.
- **AI/ML:** Artificial Intelligence and Machine Learning is increasingly being used in a number of different areas in Physical Chemistry, using techniques such as supervised learning to learn from data to make predictions, and unsupervised learning to analyse data at a higher level and find new correlations.

Internet of Things: A modular network of physical objects capable of connecting and exchanging data over the Internet.

SPARQL (SPARQL Protocol and RDF Query Language): A semantic query language for Resource Description Framework (RDF) linked data.

SQL (Structured Query Language): A query language for relational databases.

SEMANTIC WEB TECHNOLOGIES

The Semantic Web was conceptualised by Sir Tim Berners-Lee in the early 2000's with the main goal of providing a set of standards and technologies to provide machine readable and interoperable data with context and meaning (19). The Semantic Web's core technologies are Linked Data and Ontologies. Resource Description Framework (RDF) (20) is the linked data model which enables data to be broken down and represented as triples in a graph style model of the form subject → predicate → object. The predicate denotes the relationship between the subject and the object (as shown in Figure 1), and can be used to represent almost any dataset. However, RDF alone does not facilitate sufficient representation of the domain knowledge required to provide context. Ontologies, written in the Web Ontology Language (OWL) (21) enable concepts, relationships and hierarchies to be defined for the set of objects within a domain. Figure 2 shows a basic example of this with a model of a Paper, that has an author of Jane Smith. The Semantic Web also has a query language (SPARQL) (22) which works similarly to SQL but facilitates querying and traversing the graph structure of the data, enabling more complex queries to be performed.

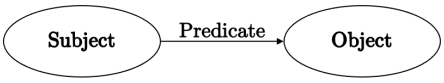

Figure 1
Linked Data Model: Abstract Model

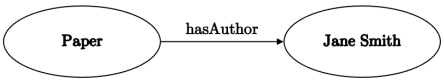

Figure 2
Linked Data Model: With Examples

Figure 3 demonstrates where the different eScience infrastructures noted above sit within the full research data life cycle. These different aspects will be discussed in further detail below, illustrating and providing examples on how they have impacted the way in which we do Physical Chemistry.

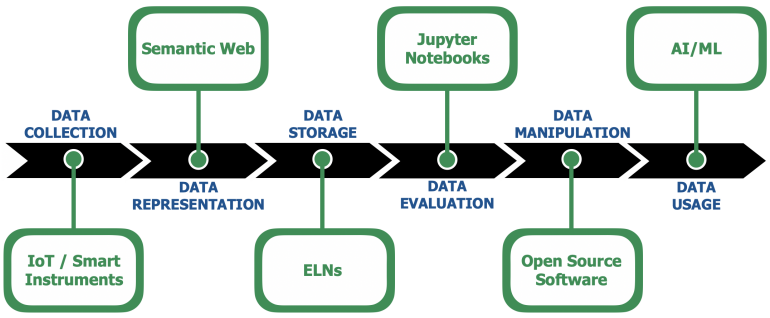

Figure 3

Demonstrating the usage of eScience infrastructures across the data lifecycle

4. COLLABORATION & THE SOCIAL MACHINE OF PHYSICAL CHEMISTRY

Research in modern physical chemistry requires a wide range of cross-disciplinary skills. A successful project often requires bringing together the expertise of chemists, data scientists, computer scientists, and equipment manufacturers together. An example of this would be bringing these interdisciplinary people together through workshops. Even a brief glance at a modern chemical laboratory (and especially a physical chemistry one) will see the importance of people, data and computers working together. These are elements of a social machine as defined by Tim Berners-Lee (23). The blend of skills required by physical chemistry researchers means they need not only chemical domain knowledge but at least a strong acquaintance with data science skills.

If, at the beginning of the 21^{st} Century, the fundamental elements of eScience and cyberinfrastructure were computationally intensive science, large datasets (later "Big Data"), and networks, the indispensability of collaboration was at least tacitly recognised alongside them. By this time, the Web was already well-established in the general scientific community as a forum for communication and the sharing of data; schemes such as Publication at Source (5) were demonstrating how methods, results, and associated research findings could be exchanged more readily. It is apparent from the evidence submitted to the panel reviewing the UK eScience programme in 2009 that collaborative research was a fundamental aspect, underpinned by powerful facilities and networked resources (1), although physical chemistry was not mentioned explicitly in that evidence.

In a talk given in 2007, Gray asserted that science, as eScience, had moved into a new pattern, relying on data exploration (24). He described this as the Fourth Paradigm, the previous three being empirical and descriptive science, theoretical science, and - in the mid-20^{th} century - computational approaches. Gray's ideas led to the publication in 2009 of The Fourth Paradigm (4), a book that, in its coverage of the many facets of data-intensive science, identifies the elements that each, in their own way, facilitate the collaborative

environments that are a key strand of eScience.

- Hunt et al [p.25] submit that "simple collaborative tools in the cloud can greatly reduce the logistics required to publish a paper" (25);
- Goble and De Roure [p.137] argue the value of computational workflows for scientific research and collaboration (26);
- Lynch [p.178] portrays the scientific record as "a vehicle for building up communities and for a form of large-scale collaboration across space and time" (27);
- Fitzgerald et al [p.204] discuss data sharing, saying that its principles "are widely acknowledged to be not only beneficial but crucial to information flows and the availability of data," although go on to examine the potential barriers (28).

Jirotka et al examined the prospects for collaborative working to enhance inter- and multi-disciplinary research (29), concluding that there was a need (in 2013) "to undertake investigations in the complexities of e-Science collaboration and the design of collaborative systems for e-Science." Also in 2013, Bird and Frey reviewed the state of information and data sharing in the chemical sciences (30). While we are now more likely to see references to "digital chemistry", it is notable that data sharing and data standards are still seen as the basis for successful collaboration (31).

Many of these practices are exemplified by the evolution of digital chemistry at Southampton, as described in (32). That paper considered three main themes: chemical information representation; changes in laboratory practice; and digital repositories. The first theme essentially explored the integration of chemical information via the Semantic Web, as described in Section 2 of this review. The work on Electronic Laboratory Notebooks (ELNs) encompassed the capture, representation, and persistent recording of experiment plans and results. Combining these aims with a focus on usability led to the LabTrove ELN (33), which has been used by several research teams, one being Todd's open notebook consortium (34). A range of digital repository activities evolved from CombeChem (11), of which the eCrystals Federation project (35) and the CrystalGrid Network (36) are the most pertinent to physical chemistry. Southampton is the primary location for the National Crystallography Service (NCS) (37).

Formal networks such as Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery (AI3SD) have enhanced interdisciplinary collaboration by forming communities of researchers who bring their specialised knowledge to bear on the challenges of scientific discovery (38).

Behaviours have generally evolved, leading to greater openness, not only in access to data, but also in the conduct of science itself. Indeed, it is possible to trace a path from eScience to Open Science (39, 32). However, replication and reproducibility remain as issues with the potential to inhibit collaborative endeavours (40, 41).

5. SMART LABORATORIES & REMOTE EXPERIMENTATION

Computer control of equipment has long been a feature of physical chemistry laboratories. Fully integrated laboratory monitoring, equipment control and data services have been developed for larger scale facilities, and as mentioned above these may be available remotely. Developing and maintaining similar environments for single labs has been a much greater challenge and, in our experience, the seemingly endless changes in sensor connections, new software, and new data formats, requires the lab systems to be recreated far too often. The

rise of the Internet of Things has led to far greater modularity in the systems and the hope that we can avoid the need to redevelop the underlying systems.

With growing modularity in software and hardware we are moving to much more interoperable and sustainable laboratory systems that enable other services to be supported on top of the well-designed lab. For example, Knight et al discuss smart laboratories, particularly the use of IoT devices, specifically for controlling experiments using x-ray radiation for imaging and spectroscopy (42) with a laboratory information architecture based on the message passing MQTT protocol (43, 44, 45).

There are challenges of scale even in a single lab. The proliferation of digital instruments and sensors has led to the majority of data being born digital, and many experiments can produce such huge quantities of data that it cannot be managed through traditional analytical techniques. The sheer power of computing that exists today means that certain manipulations that were unfeasible previously are now routine. GPU hardware opens up new possibilities for transforming observations into useful data. The challenge to curate the diversity of data produced on this scale by many research groups and centres and to make it findable and available are pushing the limits of current knowledge engineering.

Some examples of interesting projects where new technologies have been used to improve the laboratory environment:

- Example of remote monitoring and smart labs where the modularity of IoT system has greatly reduced the cost and expanded the potential uses of connected sensors (46) with a comparison of bespoke systems with what consumer level IoT can deliver.
- In a more biology focussed laboratory, the BioTISCH project replaced traditional workbench with glass covered tabletop system that presents information on the benches surface and augments objects in the form of reagent sensors. Data can be accessed from a database, calculations performed, prompts for actions, and steps in the experiment can be recorded (47). Even wearable digitisation system are becoming useable, and from the days of Google Glass, a head-mounted display for taking pictures and recording comments, with a smart watch that records motion for activity detection, and an RFID reader for use with tagged containers proved very useful (48). A related system is the Interactive tabletop system for biology lab eLabBench (49) and the Labscape in cell biology (50) in which the system directs a workflow in the lab based on a pre-defined plan and capture of the experiment keeping synchronised with the experiments through a series of connected sensors in the lab including infrared tags, barcodes and readers for tracking individual samples, tablets at each bench, pipettes monitored by cameras.
- The Ami chemistry project (chemistry) (51) uses a variety of cameras coupled with various other sensors to facilitate the monitoring and collection of experimental data in order to determine the cause of unexpected results in chemistry experiments. The monitoring of reaction conditions makes use of multi-angle video, movement sensors, and RFID tags, so all equipment and materials used can be logged and a microphone let researchers record observations.

Overall, these infrastructures make conducting scientific research in the physical chemistry domain easier. The trend to construct full Digital Twins of experiments (and not just a computer interface to each piece of equipment) will provide a vast improvement to a scientist's capacity to control and optimize experiments (52).

DIGITAL TWINS

A digital twin is a virtual computational model of a process, product, or service. The digital model can be used to analyse and predict the behaviour of the physical system. They can be continually updated, used in real time comparing the simulation with the physical process, obtaining data from, and sending data to, the physical system. They can help to prevent or diagnose issues and inform potential interventions. Digital Twins make use of data driven machine learning models as well as including science-based knowledge models of the physical system (53, 54, 55).

6. DATA COLLECTION, STORAGE & CURATION

Effective management of data is vital to the future of chemistry and for addressing the grand challenges to society. We not only need to effectively manage our own data so it is available for our own use in the future but also to make that data findable and usable by the wider scientific community; our data may be invaluable not only to physical chemistry but other disciplines as well. As mentioned previously there are a variety of data management challenges to address in order to preserve data for use in the future and make it accessible and usable by others. These challenges are common across all fields of chemistry, and indeed other sciences too. The development of effective eScience infrastructures is vital to assist with these challenges and also to improve the ways we do science. Much can be learnt about potential solutions and procedural requirements from other disciplines, especially those with a long history of collaboration and shared infrastructures such as astronomy and genetics.

Tools for data management can be considered at two different levels; those designed to assist in data management at a local or project level; and those that are designed to facilitate the sharing, analysis, and reuse of data at a community level. At the project level, initial concerns are data collection and local storage for the purposes of data analysis and future publication. The gaps between project and community data are not entirely separate, as inevitably access is required within a project to additional data for planning of experiments or comparison of results. Although this data may be institutional data from previous projects or experiments, in many cases external data is also required to build upon the results of research conducted and published within the community.

In this section we will consider the tools that address data management at the project level; and also the capabilities needed to take data into the next level and make it available and usable for the wider community.

Within the laboratory there is a need to manage the planning and recording of experiments, procedures and workflows, and manage the flow of information throughout the laboratory. Laboratory Information Management Systems (LIMS) have been used successfully in analytical laboratories for nearly thirty years, enabling the automation of laboratory tasks and facilitating the capture, storage and reporting of data and laboratory procedures (56).

Although some researchers have had reservations about using them, Electronic Laboratory Notebooks (ELNs) provide significant advances for the capture and management of the experiment records and associated data (57, 33). The capture and storage of experiments in digital note-taking tools make it much easier to locate previous experiments based upon factors such as date of experiment, materials used, sample numbers and so on. Not only do

they enable the long-term storage, enhanced access and searchability of such records, but they also facilitate curation through both user-defined and automatic generation of metadata and provenance information to describe both the notes and the associated data. Many ELNs provide valuable discipline specific tools and access to relevant databases that make the day-to-day tasks of designing and managing experiments, and the associated analyses and results much easier and more efficient for researchers. Where they are used consistently within projects ELNs have become valuable tools not just for individuals but also for groups by becoming central repositories for storing, sharing, and discussing all aspects of the research (33). Additionally, the use of standard formats and templates can improve quality and consistency in both executing and recording research (58).

For 'in silico' experimentation there are also interactive notebooks, such as Jupyter notebooks (17), that enable a researcher to capture both context and executable code in a single interactive document. A recent review of this type of software environment and how it extends into collaborative systems can be found in (59). These tools have also proved very effective for teaching programming.

These tools are effective at managing and utilising data at the level of an individual project and often provide data sharing capabilities that enable them to be shared with a wider audience, for example, for the purpose of collaboration or multi-site working. For open notebook science, some ELNs (and other note-taking tools) even allow sharing of 'live' notebooks with the wider community as a whole, such as the Open Source Malaria Notebooks (60) and the Open Lab Notebook community (61), but with the addition of Semantic Web technologies the records can be also be enriched with context and be made machine readable to facilitate their discovery and broader use (57, 62). Sharing of data before publication is still rare, but it provides the opportunity for early feedback and validation of methods and results.

Although there is still a reluctance amongst researchers to share data, even after publication, there are a number of drives under way to encourage and facilitate the practice for the benefit of science. The Fair Guiding Principles (FAIR) were developed in response to the recognition of an urgent need for infrastructure to support the reuse of scholarly data (63). FAIR refers to the characteristics of Findability, Accessibility, Interoperability, and Reusability that are required for computer systems to be able to automatically discover and reuse data. Not only does data need to be made available but it needs to be meaningful to machines. This requires a process of curation to ensure that the data is annotated with metadata, structured appropriately, and provided with a globally unique and persistent identifier, so that computer agents can identify the structure, intent and availability for reuse (16).

Curation also captures provenance information which is essential for understanding the source and quality of the data, and the process chain which may have transformed it. Although in many cases curation is an activity that takes place after the data has been published and as part of preservation in a repository, we have long advocated for active curation throughout the experiment life-cycle, in particular so that meaningful context is created at source by the researchers who best know the data (64). The process of curation can be manual and burdensome, but is significantly assisted by digitised instruments and tools such as ELNs, which support the process through automatic capture of metadata and provenance, and by avoiding the use of proprietary formats.

There is also a drive to ensure that the shared data is reproducible. This requires inclusion of important context information along with the data, such as methods, algorithms

code, and workflows, as well as the use of standard formats to ensure the data is usable and compatible with other datasets and software packages. Some authors advocate for the packaging up of these resources for deposition in repositories using standard structures and metadata to ensure that they are machine-readable and interoperable (65). The use of Data Management Plans (DMP), increasingly required by funders, has encouraged projects to consider the long-term fate of data, especially in terms of preservation and sharing (66). Many funders and other organisations such as regional data centres, disciplinary societies, and universities provide guidance or DMP templates, for example the Wellcome Trust (67) provide guidance for an outputs management plan and there is a standard template for ESRC-funded projects (68). The Digital Curation Centre (DCC) provides numerous example plans from different disciplines (69) as well as an online DMP creation tool (70) to help researchers create DMPs for their own projects. DMPs provide prompts to encourage research groups to think and record information about the nature of the data being collected, including issues around intellectual property, privacy and sensitive data; how the data will be collected including volume, structure and format of data and how these relate to commonly used formats within the field; metadata capture; data quality; how data will be organised, managed and stored; and also how it will be shared, licensed and cited; and ultimately how it will be preserved and disposed of. In addition, DMPs may also ask the researchers to consider what existing data can be reused for the project to avoid unnecessary duplication of effort and expenditure. Important technologies have been developed to facilitate the sharing of data and reproducibility of results, including repositories, providers of globally unique and persistent identifiers, tools for curation and metadata for rich annotation, and handling access controls and licenses.

7. DATA ACCESS & SHARING

Once data has been prepared for preservation and sharing, the primary tools for access and reuse are databases and repositories. These databases and repositories have become essential resources for chemists, especially for the purposes of designing experiments and comparing results. Traditionally the large-scale storage mechanisms for chemistry data have been relational databases, such as MySQL, Oracle and DB2 which use Structured Query Language (SQL) to enable users to programmatically define the structure of data, to easily add and update data, and to perform advanced search queries and reporting of the data. Although powerful, relational databases have difficulties with managing very large datasets and provide limited ability to change the structure of the data over time. NoSQL or non-relational databases in contrast are designed to work with large sets of distributed data and provide better support for scaling. They enable the storage of greater amounts of data and the design of the data structure permits flexibility over time to manage future data expansion and changes in the way that data is captured or processed over time. Traditional databases are focused on query-based retrieval but are relatively unsuitable for data exploration, much more relevant to scientific studies today (71). There are a variety of NoSQL databases available that manage different kinds of data types and structures that are suitable for different purposes, some of which are suitable for managing certain types of scientific data. Davoudian et al. provide an overview of different NoSQL database types and their uses (72), and Williams and Tkachenko detail the use of NoSQL databases and related technologies for the implementation of the Royal Society of Chemistry's ChemSpider repository and their data sharing hub (73).

The kinds of data being shared in such repositories include structure-function relationships, optical properties, excited state information, quantum chemical calculations, quantum mechanical properties, performance data, structures and properties, interaction energies, benchmark classification, crystal structure predictions, and spectra. Specialist repositories include the Biological Magnetic Resonance Data Bank (74), the Cambridge Structural Database (CSD) (75), caNanoLab (76), ChemSpider (77), CompTox Chemicals Dashboard (78), Crystallography Open Database (79), EMDataBank (80), Peptide Atlas (81), Protein Data Bank (PDB) (82), and PubChem (83).

Computational Chemistry repositories include iBIOMES for managing of large biomolecular simulation and computational chemistry data sets (84); the NIST Computational Chemistry Comparison and the Benchmark DataBase (CCCBDB) for experimental and calculated thermochemical properties (85); Benchmark Energy and Geometry Database (BEGDB) for calculations of molecular structures, energies and properties (86); and Quixote for quantum chemistry results (87). Many repositories now go beyond simple search and storage, providing tools for data creation and curation, data extraction, analysis and publication, for example ioChem-BD (88).

Materials is an area of chemistry where a lot of work has gone into the development and consideration of the requirements of FAIR databases and repositories to provide reliable sources of data for the benefit of not only individual researchers within scientific communities, but also for computational agents for applications such as machine learning. Examples include the Materials Ultimate Search Engine (MUSE) (89), the Materials Project (90), Open Quantum Materials Database (OQMD) (91), the Novel Materials Discovery (NOMAD) repository (92), Automatic Flow for materials discovery (AFLOW) (93), the ioChem-BD platform (94), and the Computational Materials Repository (CMR) (95), and Catalysis-Hub.org (96).

There are several chemistry examples of repositories used in combination with workflows and other scripts to combine and share data in a meaningful way. For example, the BioCatNet (97) database supports the discovery of enzymes by linking sequence, structure, and biochemical data from different repositories with experimental data, creating a much more complete set of information for experimental use than was previously available. The Molecular Sciences Software Institute's (MolSSI) Quantum Chemistry Archive project (98) makes use of a central server and Python infrastructure to create a community service for the automatic computation, storage, and management of quantum chemistry computations for machine learning.

The Materials Experiment and Analysis Database (MEAD) (99) automatically collects raw data from instruments during materials synthesis and characterization experiments. Nightly the data is analysed and distilled into property and performance metrics, which are added to a searchable open-source repository.

Spectroscopy is a developing area in terms of infrastructure for managing new image data and utilising existing data to support the rapid discovery of materials. For example, the SMART workflow system (100) describes the automation of computational spectroscopy to simplify data analysis and perform comparisons between theory and experiments, creating a community hub for sharing and utilising data. Highlighting the value of community adoption of standards, the Universal Spectroscopy and Imaging Data (USID) data model and Pycroscopy (101) project has defined a model that can represent any kind of imaging data from any instrument in a standard way, and facilitating access and curation for use with a multitude of platforms.

8. THE INFLUENCE OF OPEN SCIENCE

The Web has given unprecedented access to almost limitless amounts of information, generating ever higher expectations for easy access, as much for scientists as for anyone. Behaviours have shifted to recognise the need for openness in providing information as well as in accessing it. For example, Coudert makes the case that computational chemistry depends on open data, open input and output, and open software to achieve reproducible research (102).

Open Access to publications tends now to be taken for granted, perhaps disregarding the need for someone to pay the costs. Open Source software is now commonplace, recognising that community development is the best antidote to the quirks that often limited the value of 'black box' code. Sharing data for reuse and repurposing depends not only on maintaining open repositories and databases, but crucially also on the adoption of open formats.

The equipment and instruments used by physical chemists now routinely preserve the data they generate in an open format, a statement that would certainly not have been true at the outset of eScience. For example, chemical substances are identified by an International Chemical Identifier (InChI), which can be condensed to an InChIKey (103); standard open formats are available for the data produced by spectrometers and ontologies have been developed for representing physical and chemical properties and their units with established formats (87, 104, 52, 105).

Complementing the software platforms, databases, and publishing tools, open-source notebook tools have been created to aid scientists in different areas of physical chemistry research, as described in Section 6. While there are potential issues with full openness, typically intellectual property, recognition, and long-term sustainability, the nature of physical chemistry is such that these considerations are appreciably less likely to intrude than with research into novel drugs, for example. Open Data platforms can be expensive to host and to keep up to date and there are also costs involved in data valuation and curation (106), but accessible data will drive new research.

9. KNOWLEDGE REPRESENTATION & ELICITATION

As the plethora of data associated with physical chemistry and other sciences research increased, researchers looked for potential solutions to manage their data more effectively to enhance their research (107), and they settled on the Semantic Web. In the first decade, the use of Semantic Web technologies in scientific domains was mainly concentrated around the life sciences (108). As the movement started to gain popularity and scientists realised that better data management was required, more researchers in different scientific domains started to look to Semantic Web technologies to aid with their research.

In 2014, Borkum et al's paper (109) noted that the main use of Semantic Web technologies for physical chemistry was to provide controlled vocabularies and databases to semantically represent information in three areas. These were controlled vocabularies for the relevant quantities, units and symbols, and for classifying and labeling chemical substances and mixtures, and a database of chemical identifiers. However, many of the different sub domains of physical chemistry were not making use of Semantic Web technologies until recently. In the last five or so years, many more semantic resources have been created and used to advance different aspects of physical chemistry, such as McCusker et al (104) who created the NanoMine knowledge graph that contains integrated data from over 1,700 different polymer nanocomposite experiments; Farazi et al (52, 110) who created the On-

toKin ontology that describes the required domain concepts for chemical kinetic reaction mechanisms; and Phadungsukanan et al (105) who created a subdomain chemistry format, based on the Chemical Markup Language (CML) to store computational chemistry data.

Prior to 2017 there were very limited mentions of quantum chemistry in the Semantic Web sphere, though some work had been done to semantically represent quantum chemistry calculations and data (87, 105). In the last four years there has been a surge of activity in this area, with ontologies, and semantic platforms being created for use in quantum chemistry (111, 112, 110). A notable contribution in this field is that of Krdzavac et al (113), who created the OntoCompChem ontology to semantically represent quantum chemical calculations. This has been used alongside knowledge graphs and software agents to advance other areas of physical chemistry such as thermochemistry (113) and reaction kinetics (110). Wang et al (114) created a computational chemistry data management platform that uses ontology-based methods of thermophysical data integration (115) and looks to ensure that researchers can share and reuse each other's data in a consistent, comparable, interoperable format.

Work has also been conducted in the materials space to create knowledge graphs and ontologies for materials science (104, 116) to enable scientists working in this area to integrate their data and visualise their datasets in different ways. These technologies have also been used in conjunction with machine learning technologies. Picklum and Beetz (117) recently worked on knowledge-enabled machine learning in the materials sciences, using modern machine learning techniques with machine-readable semantic datasets to exploit the links between the different datasets and enable complex queries and reasoning to be performed.

New ontologies have also been created to capture the data and semantics of chemical kinetic reaction mechanisms. A notable ontology in this area is OntoKin, developed by Farazi et al (52), which was created to enable researchers to query, compare, and retrieve mechanisms via the Semantic Web. This ontology is used in a knowledge graph that addresses inconsistency issues in chemical mechanisms (118).

The field of spectroscopy has also made recent headway in the use of Semantic Web technologies. Whilst it is not a new realisation that the creation of metadata and ontologies is important for the field of spectroscopy (119, 120), the actual development of ontologies and models, particularly to support processing tabular and graphical resources in qualitative spectroscopy, has occurred only in the last few years (121).

Overall, there has been considerable work in the last five years in many subdomains of physical chemistry to create ontologies to ensure that data can be represented and interlinked in a consistent manner, in addition to creating various platforms to facilitate data sharing and integration. This is as much a human endeavour as a technological one and is both a digital and collaboration-based infrastructure of eScience.

10. PERSPECTIVE

eScience as a name has somewhat disappeared. Precisely because it has been so successful, its achievements have become part of the expected scientific infrastructure, and physical chemistry has certainly benefited from these advances. While the topics and fundamental ideas involved in experimental physical chemistry research have not been altered in the last two decades by advances in digital infrastructures, these advances and indeed the emergence of the fourth industrial revolution (Industry 4.0) (122), have made a significant impact on

how we are able to actually conduct physical chemistry research.

We are moving into an ever increasing digital era. There are now technologies and infrastructures to support doing physical chemistry at every stage of the research lifecycle. How we collect, curate, store and make our data available underpins every aspect of future research based on that data, and managing our data well is essential for facilitating and enhancing knowledge discovery and innovation.

The recently published book Data Science in Chemistry (123) provides an encyclopaedic survey of the data science techniques and methodologies that can be applied to physical chemistry, and indeed to all other branches of chemistry. It shows that Data Science has permeated physical chemistry, making it easier for physical chemists to successfully apply the range of AI/ML technologies that are now becoming available. It is also worth noting that collaboration has played a key part in enabling all these infrastructures.

Creating open source software and databases, making open controlled vocabularies for reuse, and providing algorithms and programming libraries, are all born out of collaboration. This is reflected in the fact that these infrastructures and the advances in conducting science that they afford would not be possible without human effort; the combined endeavours of the open physical science community have done as much as the technology itself.

Git: A command line based version control system that manages and stores revisions of projects.

GitHub: An open source project that hosts Git repositories and provides a web-based graphical interface as an alternative to the command line.

We are still very much in a liminal (transition) period, but changes are being driven by both open science and compliance with the data management objectives of the Research Councils. It is vital that data and methods of publications are available to anyone who might desire them. The use of computational notebooks (such as Jupyter notebooks) for data analysis in experimental and computational physical chemistry is one example that facilitates efficient dissemination. The explanation, code and data can all be readily integrated and shared for example via the version control of software sharing site Git/GitHub (124) to facilitate collaboration and the use and reuse of software. Many journals now require not only the data to be available alongside a paper but the software code as well, though ensuring that these can be readily reused is still a problem with many publication formats.

We see eScience as very much an enabler and facilitator of open science both in terms of development of the technology and the changes in viewpoint of human and computer interactions – a true example of a successful social machine.

We would claim that many of the advances enabled by eScience led us to be able to adapt to the COVID-19 situation far more rapidly and flexibly. For example, scientific collaboration via video conferencing enabled global interdisciplinary teams to coordinate the collaboration between physical chemists with medicinal chemists and other experts. The exchange of even large amounts of data in computer readable form, using internationally agreed standards has been enormously facilitated by the eScience infrastructures (even if there is still some way to go in achieving the FAIR ideals). The remote operation of experiments, and remote support for researchers, enabled many laboratories to continue to run effectively even during lock-down conditions.

Ultimately, the process of "doing the science" is being made much easier. The new infrastructures have brought more immediate computational power to the experimentalists, enabling them to make the best of their data and provided a better working environment for our researchers in physical chemistry

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank our many colleagues and students from Chemistry, Physics, Computer Science, Mathematics & Statistics, and the Humanities, from the University, Industry and Government sectors who have contributed to our understanding of eScience and the development of our own work in this area. Our work on eScience and Digital Economy has been funded by the UK EPSRC under grants GR/R67729/01 Structure-Property Mapping: Combination Chemistry & the grid (CombeChem), EP/C008863/1 End-to-End pipeline for chemical information: from the laboratory to literature and back, EP/K003569/ Digital Economy IT as a Utility Network$^+$, EP/N014189/1 Joining the dots: from data to insight, EP/S000356/1 Artificial and Augmented Intelligence for Automated Scientific Discovery

LITERATURE CITED

1. Atkins DE, Borgman CL, Bindhoff N, Ellisman M, Felman S, et al. 2010. RCUK Review of e-Science 2009: Building a UK foundation for the transformative enhancement of research innovation. Tech. rep., RCUK
2. Wikipedia. 2021. e-Science. `https://en.wikipedia.org/w/index.php?title=E-Science&oldid=1016428593`
3. IGI. 2021. What is E-Science. `https://www.igi-global.com/dictionary/provenance-tracking-end-user-oriented/8912`
4. Hey T, Tansley S, Tolle K. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft
5. Frey JG, Roure DD, Carr L. 2002. Publication at source: scientific communication from a publication web to a data grid. In *EuroWeb 2002 Conference*
6. Hey T, Trefethen AE. 2002. The UK e-Science Core Programme and the Grid. In *Computational Science — ICCS 2002*, eds. PMA Sloot, AG Hoekstra, CJK Tan, JJ Dongarra, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer
7. Gervasi O, Dittamo C, Laganà A. 2005. A Grid Molecular Simulator for E-Science. In *Advances in Grid Computing - EGC 2005*, eds. PMA Sloot, AG Hoekstra, T Priol, A Reinefeld, M Bubak, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer
8. Robinson JM, Frey JG, Stanford-Clark AJ, Reynolds AD, Bedi BV. 2005. Sensor networks and grid middleware for laboratory monitoring. In *First International Conference on e-Science and Grid Computing (e-Science'05)*. IEEE
9. Wang J, Korambath P, Kim S, Johnson S, Jin K, et al. 2011. Facilitating e-Science Discovery Using Scientific Workflows on the Grid. In *Guide to e-Science: Next Generation Scientific Research and Discovery*, eds. X Yang, L Wang, W Jie, Computer Communications and Networks. London: Springer, 353–382
10. Khillar S. 2018. Difference between Grid Computing and Cloud Computing. `http://www.differencebetween.net/technology/difference-between-grid-computing-and-cloud-computing/`
11. Frey J, De Roure D, Taylor K, Essex J, Mills H, Zaluska E. 2006. Combechem: a case study in provenance and annotation using the semantic web. In *International Provenance and Annotation Workshop*. Springer

12. Taylor KR, Essex JW, Frey JG, Mills HR, Hughes G, Zaluska E. 2006. The semantic grid and chemistry: experiences with combechem. *Journal of Web Semantics* 4:84–101
13. Hey T. 2020. AI3SD Video: AI for Science: Transforming Scientific Research. In *AI3SD Summer Seminar Series*, eds. S Kanza, JG Frey, M Niranjan, V Hooper
14. Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* 115:2584–2589. Publisher: National Academy of Sciences Section: Colloquium Paper
15. Halford S, Pope C, Carr L. 2010. A manifesto for Web Science. In *WebSci10: Extending the Frontiers of Society On-Line*, eds. J Erickson, S Gradmann. Raleigh, United States
16. Bird CL, Willoughby C, Coles SJ, Frey JG. 2013. Data Curation Issues in the Chemical Sciences. *Information Standards Quarterly* 25:4
17. ProjectJupyter. 2021. Jupyter. `https://www.jupyter.org`
18. Overleaf. 2021. Overleaf, Online LaTeX Editor. `https://www.overleaf.com`
19. Berners-Lee T, Hendler J, Lassila O. 2001. The semantic web. *Scientific american* 284:34–43
20. 2014. RDF 1.1 Concepts and Abstract Syntax. `https://www.w3.org/TR/rdf11-concepts/`
21. 2012. OWL 2 Web Ontology Language Document Overview (Second Edition). `https://www.w3.org/TR/owl2-overview/`
22. ???? SPARQL 1.1 Query Language
23. Hendler J, Berners-Lee T. 2010. From the semantic web to social machines: A research challenge for ai on the world wide web. *Artificial intelligence* 174:156–161
24. Gray J, Szalay A. 2007. escience-a transformed scientific method. *presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA*
25. Hunt JR, Baldocchi DD, van Ingen C. 2009. Redefining ecological science using data. In *The Fourth Paradigm: Data Intensive Scientific Discovery*, eds. T Hey, S Tansley, K Tolle. Microsoft Research, 21–26
26. Goble C, De Roure D. 2009. The Impact of Workflow Tools on Data-centric Research. In *The Fourth Paradigm: Data Intensive Scientific Discovery*, eds. T Hey, S Tansley, K Tolle. Microsoft Research, 137–146
27. Lynch C. 2009. Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In *The Fourth Paradigm: Data Intensive Scientific Discovery*, eds. T Hey, S Tansley, K Tolle. Microsoft Research, 177–184
28. Fitzgerald A, Fitzgerald B, Pappalardo K. 2009. The Future of Data Policy. In *The Fourth Paradigm: Data Intensive Scientific Discovery*, eds. T Hey, S Tansley, K Tolle. Microsoft Research, 201–208
29. Jirotka M, Lee CP, Olson GM. 2013. Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW)* 22:667–715
30. Bird CL, Frey JG. 2013. Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences. *Chemical Society Reviews* 42:6754–6776. Publisher: Royal Society of Chemistry
31. RSC. 2020. Welcoming a new era: A time for digital scientific discovery. `https://www.chemistryworld.com/rsc/welcoming-a-new-era-a-time-for-digital-scientific-discovery/4012131.article`
32. Bird C, Coles SJ, Frey JG. 2015. The Evolution of Digital Chemistry at Southampton. *Molecular Informatics* 34:585–597. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201500008
33. Badiola KA, Bird C, Brocklesby WS, Casson J, Chapman RT, et al. 2015. Experiences with a researcher-centric eln. *Chemical science* 6:1614–1629
34. Badiola KA, Quan DH, Triccas JA, Todd MH. 2014. Efficient synthesis and anti-tubercular activity of a series of spirocycles: an exercise in open science. *PloS one* 9:e111782
35. Lyon L, Coles S, Duke M, Koch T. 2008. Scaling up: towards a federation of crystallography data repositories

36. Coles S, Frey J, DeRoure D, Hursthouse M. 2004. The crystalgrid collaboratory foundation workshop, southampton, 13-17 september, 2004: a selection of presentations
37. 2021. UK National Crystallography Service. `http://www.ncs.ac.uk/`
38. Kanza S, Bird CL, Niranjan M, McNeill W, Frey JG. 2021. The AI for Scientific Discovery Network+. *Patterns* 2:100162
39. Frey JG, Bird C. 2018. Reducing uncertainty: the raison d'être of open science? *Beilstein Magazine* 4
40. Serra-Garcia M, Gneezy U. 2021. Nonreplicable publications are cited more than replicable ones. *Science Advances* 7:eabd1705. Publisher: American Association for the Advancement of Science Section: Research Article
41. Coveney PV, Highfield RR. 2021. When we can trust computers (and when we can't). *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379:20200067. Publisher: Royal Society
42. Knight NJ, Kanza S, Cruickshank D, Brocklesby WS, Frey JG. 2020. Talk2Lab: The Smart Lab of the Future. *IEEE Internet of Things Journal* 7:8631–8640. Conference Name: IEEE Internet of Things Journal
43. MQQT.org. 2020. MQTT: The Standard for IoT Messaging
44. Atmoko R, Riantini R, Hasin M. 2017. Iot real time data acquisition using mqtt protocol. In *Journal of Physics: Conference Series*, vol. 853, no. 1. IOP Publishing
45. Porr M, Schwarz S, Lange F, Niemeyer L, Hentrop T, et al. 2020. Bringing iot to the lab: Sila2 and open-source-powered gateway module for integrating legacy devices into the digital laboratory. *HardwareX* 8:e00118
46. Perkel JM. 2017. The internet of things comes to the lab. *Nature* 542:125–126
47. Echtler F, Häussler M, Klinker G. 2010. Biotisch :3439–3444
48. Scholl PM, Van Laerhoven K. 2014. Wearable digitization of life science experiments. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct. New York, NY, USA: Association for Computing Machinery
49. Tabard A, Hincapié-Ramos JD, Esbensen M, Bardram JE. 2011. The elabbench: An interactive tabletop system for the biology laboratory. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ITS '11. New York, NY, USA: Association for Computing Machinery
50. Arnstein L, Hung CY, Franza R, Zhou QH, Borriello G, et al. 2002. Labscape: a smart environment for the cell biology laboratory. *IEEE Pervasive Computing* 1:13–21
51. 2011. Ami - the chemist's amanuensis. *Journal of Cheminformatics* 3:45
52. Farazi F, Akroyd J, Mosbach S, Buerger P, Nurkowski D, et al. 2020. OntoKin: An Ontology for Chemical Kinetic Reaction Mechanisms. *Journal of Chemical Information and Modeling* 60:108–120. Publisher: American Chemical Society
53. Niederer SA, Sacks MS, Girolami M, Willcox K. 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science* 1:313–320
54. Armstrong MM. 2020. Cheat sheet: What is digital twin?
55. Girolami M. 2021. Digital twin technology a 'powerful tool' but requires significant investment, say experts
56. Gibbon GA. 1996. A brief history of LIMS. *Laboratory Automation & Information Management* 32:1–5
57. Kanza S, Willoughby C, Gibbins N, Whitby R, Frey JG, et al. 2017. Electronic lab notebooks: can they replace paper? *Journal of cheminformatics* 9:1–15
58. Piccione PM. 2020. Systematizing scientific laboratory work by a workflow and template for electronic laboratory notebooks. *Education for Chemical Engineers* 31:42–53
59. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, et al. 2021. Best practices in machine learning for chemistry. *Nature Chemistry* 13:505–508. Number: 6 Publisher: Nature Publishing

Group
60. OSM. 2020. OSM - Open Source Malaria. http://opensourcemalaria.org/
61. openlabnotebooks.org. 2021. Open lab notebooks. openlabnotebooks.org
62. Kanza S, Gibbins N, Frey JG. 2019. Too many tags spoil the metadata: investigating the knowledge management of scientific research with semantic web technologies. *Journal of cheminformatics* 11:1–23
63. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. Number: 1 Publisher: Nature Publishing Group
64. Frey J, Coles SJ, Bird C, Willoughby C. 2015. Collection, Curation, Citation at Source: Publication@Source 10 Years On. *International Journal of Digital Curation* 10:1–11. Number: 2
65. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I. 2010. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings* :1–1Publisher: Nature Publishing Group
66. DCC. 2021. Data Management Plans. https://www.dcc.ac.uk/resources/data-management-plans
67. Wellcome. 2014. How to complete an outputs management plan. https://wellcome.org/grant-funding/guidance/how-complete-outputs-management-plan
68. UKDS. 2014. Data management planning for ESRC researchers. https://www.ukdataservice.ac.uk/manage-data/plan/dmp-esrc
69. DCC. 2014. Example DMPs and guidance. https://dcc.ac.uk/resources/data-management-plans/guidance-examples
70. DCC. 2014. Plan to make data work for you. https://dmponline.dcc.ac.uk/
71. Fox P, Hendler J. 2011. Changing the equation on scientific data visualization. *Science* 331:705–708
72. Davoudian A, Chen L, Liu M. 2018. A survey on nosql stores. *ACM Computing Surveys (CSUR)* 51:1–43
73. Williams A, Tkachenko V. 2014. The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. *Journal of Computer-Aided Molecular Design* 28:1023–1030
74. PDB. 2021. BMRB - Biological Magnetic Resonance Bank. https://bmrb.io/
75. CCDC. 2021. The Cambridge Structural Database (CSD
76. NIH. 2021. caNanoLab. https://cananolab.nci.nih.gov/caNanoLab/#/
77. RSC. 2021. ChemSpider. http://www.chemspider.com/
78. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, et al. 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* 9:61
79. 2021. Crystallography Open Database. http://www.crystallography.net/cod/
80. 2021. The Electron Microscopy Data Bank. https://www.ebi.ac.uk/pdbe/emdb/
81. 2021. PeptideAtlas. http://www.peptideatlas.org/
82. wwPDB consortium, Burley SK, Berman HM, Bhikadiya C, Bi C, et al. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* 47:D520–D528
83. NIH. 2021. PubChem. https://pubchem.ncbi.nlm.nih.gov/
84. Thibault JC, Facelli JC, Cheatham TE. 2013. iBIOMES: Managing and Sharing Biomolecular Simulation Data in a Distributed Environment. *Journal of Chemical Information and Modeling* 53:726–736. Publisher: American Chemical Society
85. Johnson R. 2002. Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database 101. Type: dataset
86. ioChem BD. 2021. Benchmark Energy and Geometry Datanba. http://www.begdb.org/

87. Adams S, de Castro P, Echenique P, Estrada J, Hanwell MD, et al. 2011. The Quixote project: Collaborative and Open Quantum Chemistry data management in the Internet age. *Journal of Cheminformatics* 3:38
88. ioChem BD. 2021. ioChem-BD. `www.iochem-bd.org`
89. Himanen L, Geurts A, Foster AS, Rinke P. 2019. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* 6:1900808. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.201900808
90. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* 1:011002. Publisher: American Institute of Physics
91. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, et al. 2015. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* 1:1–15. Number: 1 Publisher: Nature Publishing Group
92. Draxl C, Scheffler M. 2018. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin* 43:676–682. Publisher: Cambridge University Press
93. Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, et al. 2012. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science* 58:218–226
94. Álvarez Moreno M, de Graaf C, López N, Maseras F, Poblet JM, Bo C. 2015. Managing the computational chemistry big data problem: the ioChem-BD platform. *Journal of Chemical Information and Modeling* 55:95–103
95. Landis DD, Hummelshøj JS, Nestorov S, Greeley J, Dulak M, et al. 2012. The Computational Materials Repository. *Computing in Science & Engineering* 14:51–57. Publisher: American Institute of Physics
96. Winther KT, Hoffmann MJ, Boes JR, Mamun O, Bajdich M, Bligaard T. 2019. Catalysis-Hub.org, an open electronic structure database for surface reactions. *Scientific Data* 6:75. Number: 1 Publisher: Nature Publishing Group
97. Pc B, C V, W R, M P, D R, et al. 2016. BioCatNet: A Database System for the Integration of Enzyme Sequences and Biocatalytic Experiments. *Chembiochem : a European Journal of Chemical Biology* 17:2093–2098
98. Smith DGA, Altarawy D, Burns LA, Welborn M, Naden LN, et al. 2021. The MolSSI QCArchive project: An open-source platform to compute, organize, and share quantum chemistry data. *WIREs Computational Molecular Science* 11:e1491. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1491
99. Soedarmadji E, Stein HS, Suram SK, Guevarra D, Gregoire JM. 2019. Tracking materials science data lineage to manage millions of materials experiments and analyses. *npj Computational Materials* 5:1–9. Number: 1 Publisher: Nature Publishing Group
100. Licari D, Fusè M, Salvadori A, Tasinato N, Mendolicchio M, et al. 2018. Towards the SMART workflow system for computational spectroscopy. *Physical Chemistry Chemical Physics* 20:26034–26052. Publisher: The Royal Society of Chemistry
101. Somnath S, Smith CR, Laanait N, Vasudevan RK, Jesse S. 2019. USID and Pycroscopy – Open Source Frameworks for Storing and Analyzing Imaging and Spectroscopy Data. *Microscopy and Microanalysis* 25:220–221. Publisher: Cambridge University Press
102. Coudert FX. 2017. Reproducible Research in Computational Chemistry of Materials. *Chemistry of Materials* 29:2615–2617. Publisher: American Chemical Society
103. 2021. The IUPAC International Chemical Identifier (InChI). Https://iupac.org/who-we-are/divisions/division-details/inchi/
104. McCusker JP, Keshan N, Rashid S, Deagen M, Brinson C, McGuinness DL. 2020. NanoMine: A Knowledge Graph for Nanocomposite Materials Science. In *The Semantic Web – ISWC 2020*, eds. JZ Pan, V Tamma, C d'Amato, K Janowicz, B Fu, A Polleres, O Seneviratne, L Kagal, Lecture Notes in Computer Science. Cham: Springer International Publishing

105. Phadungsukanan W, Kraft M, Townsend JA, Murray-Rust P. 2012. The semantics of Chemical Markup Language (CML) for computational chemistry : CompChem. *Journal of Cheminformatics* 4:15
106. Editorial. 2017. Empty rhetoric over data sharing slows science. *Nature* 546:327
107. Menon A, Krdzavac NB, Kraft M. 2019. From database to knowledge graph — using data in chemistry. *Current Opinion in Chemical Engineering* 26:33–37
108. 2011. Linked Data: Evolving the Web into a Global Data Space. `http://linkeddatabook.com/editions/1.0/`
109. Borkum MI, Frey JG. 2014. Usage and applications of Semantic Web techniques and technologies to support chemistry research. *Journal of Cheminformatics* 6:18
110. Farazi F, Krdzavac NB, Akroyd J, Mosbach S, Menon A, et al. 2020. Linking reaction mechanisms and quantum chemistry: An ontological approach. *Computers & Chemical Engineering* 137:106813
111. Wang B, Dobosh PA, Chalk S, Ito K, Sopek M, Ostlund NS. 2018. A Portal for Quantum Chemistry Data Based on the Semantic Web. In *Concepts, Methods and Applications of Quantum Systems in Chemistry and Physics*, eds. YA Wang, M Thachuk, R Krems, J Maruani, Progress in Theoretical Chemistry and Physics. Cham: Springer International Publishing
112. Krdzavac N, Mosbach S, Nurkowski D, Buerger P, Akroyd J, et al. 2019. An Ontology and Semantic Web Service for Quantum Chemistry Calculations. *Journal of Chemical Information and Modeling* 59:3154–3165. Publisher: American Chemical Society
113. Krdzavac N, Mosbach S, Nurkowski D, Buerger P, Akroyd J, et al. 2019. An ontology and semantic web service for quantum chemistry calculations. *Journal of chemical information and modeling* 59:3154–3165
114. Wang B, Dobosh PA, Chalk S, Sopek M, Ostlund NS. 2017. Computational Chemistry Data Management Platform Based on the Semantic Web. *The Journal of Physical Chemistry A* 121:298–307. Publisher: American Chemical Society
115. Kosinov AV, Erkimbaev AO, Zitserman VY, Kobzev GA. 2019. Ontology-based methods of thermophysical data integration. *Journal of Physics: Conference Series* 1385:012033. Publisher: IOP Publishing
116. Horsch M, Chiacchiera S, Schembera B, Seaton M, Todorov I. 2021. Semantic interoperability based on the European Materials and Modelling Ontology and its ontological paradigm: Mereosemiotics. In *Proceedings of WCCM-ECCOMAS 2020*
117. Picklum M, Beetz M. 2019. MatCALO: Knowledge-enabled machine learning in materials science. *Computational Materials Science* 163:50–62
118. Farazi F, Salamanca M, Mosbach S, Akroyd J, Eibeck A, et al. 2020. Knowledge Graph Approach to Combustion Chemistry and Interoperability. *ACS Omega* 5:18342–18348. Publisher: American Chemical Society
119. Lavrentiev N, Privezentsev A, Fazliev A. 2008. Informational system for the solution of molecular spectroscopy problems. 4. Transitions in molecules of C2v and Cs symmetry. *Atmospheric and oceanic optics* 21:836–841
120. Lavrentiev NA, Rodimova OB, Fazliev AZ, Vigasin AA. 2017. Systematization of published research graphics characterizing weakly bound molecular complexes with carbon dioxide. In *23rd International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics*, vol. 10466. International Society for Optics and Photonics
121. Lavrentiev NA, Privezentsev AI, Fazliev AZ. 2019. Tabular and Graphic Resources in Quantitative Spectroscopy. In *Data Analytics and Management in Data Intensive Domains*, eds. Y Manolopoulos, S Stupnikov, Communications in Computer and Information Science. Cham: Springer International Publishing
122. Schwab K. 2017. The fourth industrial revolution. Currency
123. Gressling T. 2020. Data Science in Chemistry. De Gruyter. Publication Title: Data Science in Chemistry

124. GitHub. 2022. Github