
Supplementary information

**Making the collective knowledge of
chemistry open and machine actionable**

In the format provided by the
authors and unedited

**Supplementary Information for:
Making the collective knowledge of chemistry open and
machine actionable**

Kevin Maik Jablonka¹, Luc Patiny²✉, Berend Smit¹✉

¹Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingenierie Chimiques (ISIC), École Polytechnique Fédérale de Lausanne (EPFL), Sion, VS, Switzerland.

²Institut des Sciences et Ingénierie Chimiques (ISIC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

✉ luc.patiny@epfl.ch, ✉ berend.smit@epfl.ch,

Contents

1	Implementing and Maintaining an Open Science Infrastructure	3
1.1	Survival advantage of the worse	3
1.2	Software can be maintained and improved by the community	4
1.3	Modularity, on different levels, is key	5
1.4	The user base is heterogeneous	6
1.5	Convergence and interoperability	7

Supplementary Note 1 Implementing and Maintaining an Open Science Infrastructure

In this part, we want to focus on the implementation, discussing some key lessons we learned from implementing an electronic lab notebook (ELN). We think that chemistry can learn several lessons from the world of open source.

1.1 Survival advantage of the worse

In a networked world, the worse solution can have an advantage over an over-engineered one (the infamous “worse is better” principle).¹ Over the years, the chemistry community could observe the proposal of a welter of ontologies, schema, and ELNs. While those are admirable efforts we feel that new file formats will not help the community at this point. In particular, they will not address the biggest problem which is the lack of interoperability between existing solutions. For most data types there are already standardised, sometimes even IUPAC recommended vocabularies, schema, and serialization formats, and from our experience one cannot anticipate all eventualities of a data schema (chemistry is likely just too flexible, some even argue that a schema-first approach will never scale²). Many successful technologies experienced that after an initial design (that is good enough for a potential disruptive innovation) it is much more worthwhile to implement a prototype and keep on iterating based on user feedback (which will be conditioned on an imperfect design, but stick to it as it provides them value). The “worse” solution will have better survival characteristics because it is easier to iterate on it (whereas an overly complicated solution might become impossible to maintain).

In this context, it is also interesting to reflect on what Oleksik et al.³ described as “a tension that is intrinsic to the digital nature of ELNs: a conflict between the flexibility, fluidity, and low threshold for modifying digital records and the requirement for persistence and consistency”. In the ELNs that are currently used in the chemistry community, one finds both extremes: Some interfaces are reminiscent of filling a long form (like for a tax declaration), whereas others provide no structure at all (like a simple note-taking app). From the viewpoint of data capture (seeing the data analysts as the end-users) a highly structured form might be “best”—but it will be a worse overall solution as chemists who are supposed to enter the data will not adopt it.

We find that a key design requirement is to make the barrier for data entry as low as possible, potentially using some ideas like predefined sentences with variable fields, to increase adoption. For chemistry, this also means that an ELNs should support the editing of chemical structures—also to be able to store this information in a reusable form and not as an image that was created by another software.

1.2 Software can be maintained and improved by the community

Some of our experimental colleagues raised the viewpoint that complicated pieces of software cannot be maintained by researchers.

Our everyday life shows that the opposite is true. We all use software that is developed and maintained by the community. Our world would, as we know it, would not be possible without open-source projects such as Linux, Firefox, or Python. All these pieces of software are probably better because they are open, hence better tested (“given enough eyeballs, all bugs are shallow”⁴). Clearly, there are issues with the sustainability of open-source software, and there are initiatives that call for novel funding schemes^{5,6}—but there are also many examples that show that successful businesses can be built around open software. For example, the SciNote ELN is licensed under the open Mozilla public license, but there is still a business model that sells deployments of the ELN on the cloud together with support—for users that do not want to deploy the ELN themselves. A similar scheme is used for the eLabFTW ELN where Deltablott sells “pro support” and hosting, i.e. service-level agreements (SLAs).

On the other hand, we can think of an open science infrastructure as something like our NMR facility. Our departments have funds to keep this infrastructure running. To us, it seems forward-looking to do the same for a data infrastructure that not only supports research but could also enable new research.

Notably, the open-source model has not only been proven successful for software development but also for the sciences. One example is the polymath project where the maths professor Timothy Gowers used his blog to ask for help with a proof for a central theorem.⁷ Within less than two months, the results—based on about 800 comments of 27 people—were being written up in a paper. Similarly, an open-source science project found a new pathway to an enantiopure form of the drug praziquantel.⁸ One of the reasons why this model works is that the relevant expertise and the new viewpoints can make themselves heard—no research group can have all the experience and expertise there is in the world, and often an overly

specialised team can struggle to escape a local minimum in their thinking.^{8,9} The acceleration of research that both the polymath and the open-source malaria project observed was described by Timothy Gowers, the initiator of the polymath project, as “It feels as though this process is to normal research as driving is to pushing a car.”¹⁰

As with open-source software, open-source science is also interesting because the process of creation is open. We can follow all the discussions that led to the development of Linux, we can follow all the small steps that have been made for the polymath project, and can also see how the route to the enantiopure form of praziquantel was discovered. All this can be used as a resource for education, but also to understand why things are as they are. An open science infrastructure such as the one outlined in this work can be a key ingredient for this process—by simply opening the lab notebooks and allowing others to comment on it. One can even envision that (parts of) the data captured via the ELNs is fed directly as a dataset into the Kaggle platform,¹¹ where (aspiring) data scientists can use it to develop new machine learning models.

1.3 Modularity, on different levels, is key

An ELN is a complex piece of software. If one were to develop it as one big monolith it would be infeasible to try to fix or update one part of the code without breaking another one, and it would also hamper parallel development on different parts of the codebase by different developers (“Linus does not scale”). A monolith is not scalable. Simply alone for this reason it is key that the tool is developed at the most granular level. But besides making development and maintenance easier it comes with other advantages. The world of (web) development is fast changing. No one knows what the new, state-of-the-art technologies will be in 10 years. By tightly coupling all parts of the ELN it will be completely unfeasible to upgrade to a new framework. For example, if one keeps the frontend (the client side code that runs in the browser and visualises the data) separate from the backend (the code that runs on the server and does the heavy lifting) one can much easier migrate to new frameworks.¹² One new ELN framework that followed this approach is the Chemotion ELN developed at the Karlsruhe Institute for Technology (KIT). One advantage of developing the frontend in a modern web framework is that the resulting service can be used on any platform—also on mobile devices—via the web browser.

One challenge digital records face compared to paper-based notes is immediacy.¹³ It is usually much faster to quickly jot something down on paper

than to enter it via a clumsy interface on a computer that might not even be in the lab. For this reason, we envision that the ELN must improve the ways researchers can ingest data. One way we explored in our work are chatbots¹⁴ that easily allow capturing pictures or videos—something that is typically much harder to do with paper-based notebooks and that would allow capturing the experiments in much more detail. Others explored to improve the ease of use using speech recognition,^{15;16} but these techniques did not find widespread use, one reason for which the specialised chemistry vocabulary might be.

The LabTrove ELN showcases an advantage digital solutions can have over the conventional paper-based notebooks: They allow to connect the reflection about the data directly with the data records themselves. In the context of LabTrove one might write summary blog posts in preparation for weekly group seminars that link to particular experiments. In this way, not only the experiments themselves but also the reflection on the experiments and the thought processes that led to the next steps, are captured—similar to what happens in open source science projects.

1.4 The user base is heterogeneous

One challenge every ELN faces is that the user base is heterogeneous. On the one hand, this stems from the fact that, in our perspective, the users are not only the experimental chemists but also scientists that will re-purpose the data. Both favour different ways of interacting with the system. But also within the group of experimental chemists, there are varying amounts of interest in what is going on behind the scenes. On top of that, chemists will attempt to use the ELN from different platforms. Many would prefer to use it from a mobile device, but access also from their Mac, Windows, or Linux desktop computer. One viable solution to address all these different needs is to offer the ELN as a cloud service that can be accessed from a web browser on any device. This also simplifies the rollout of fixes and updates as the developers do not depend on users installing them. Lastly, this also simplifies the life of users as they do not need to install anything—they just need to open their web browsers. Note that it can make sense to implement the ELN such that most operations take place on the client-side. This is feasible as the most common operations in the ELN are not resource-intensive (e.g. looking up data in a database, plotting spectra, writing notes). It can also make the ELN more responsive than design in which *all operations* take place on a server (e.g., implemented using Jupyter notebooks, where

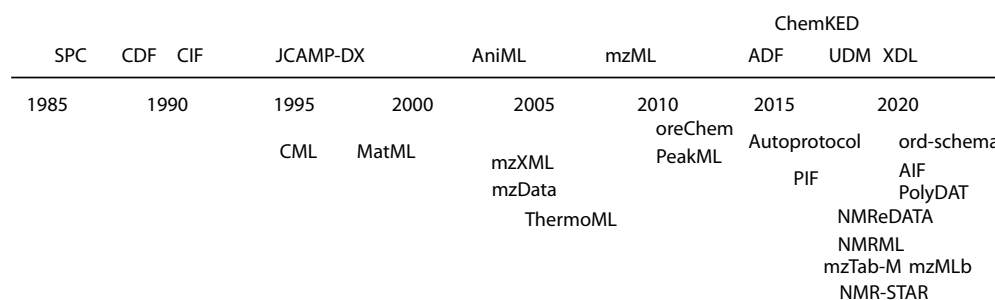
one would need to start one container per user and cannot profit from many of the optimisation techniques that modern web frameworks offer).

1.5 Convergence and interoperability

For this vision to be sustainable on a larger scale different solutions must be able to talk to each other. It is clear that there cannot be only one database or only one ELNs as the needs of different communities are just too diverse. But the different tools need to be interoperable to make data reusable and harvestable on a larger scale. Standardisation will also be key to integrate “manual chemistry” captured via ELNs with syntheses executed by robots and to link experimental work with computational works.

Supplementary Table 1 | Some examples (non-comprehensive) of schema that have been developed for chemistry and materials science. A more comprehensive overview can be found on <https://fairsharing.org/> via search for “standards”. There have been many efforts that did not find widespread use.

schema	description	
Analytical Markup Language (An-IML) ¹⁷	Information	focused on analytical chemistry and biological data, developed by an ASTM (American Society for Testing and Materials) working group. Found already support by some manufacturers like Agilent or ELNs like LabWare. BSSN software (now owned by Merck) implemented converters for more than 150 instruments ¹⁸
autoprotocol ¹⁹		development focused on defining experimental plans in the life sciences that can then be remotely executed, ¹⁹ but has elements that are general (e.g., compounds, inventory) and cover different aspects of chemistry. Notably, it has a mechanism for the proposal of changes to the standard (Autoprotocol Standard Changes)
Physical Information File (PIF) ²⁰	Information File	designed to be able to describe a broad array of data “from parts in a car down to a single monomer in a polymer matrix”, hence can have a nesting of subsystems. Maintained by Citrine Informatics
Graphical Expression of Materials Data		Developed by Citrine Informatics, links materials, the processes that produce them, and the measurements that characterize them. It resolves some problems that were present in the data model underlying the PIF
ThermoML ²¹		IUPAC standard for storage and exchange of experimental thermo-physical and thermochemical property data. Data produced by the NIST Thermodynamics Research Center are provided in this format
Chemical Markup Language (CML) ^{22;23}	Language	is supported by many common programs and has been extended, for example, for spectra (CMLSpect) ²⁴ . One core idea of designers of the CML was that the information should be validatable—for example, to automatically find if there negative atom counts
schema for the Open Reaction Database (ord-schema)		the developers built this schema based on a survey they conducted in late 2019/early 2020 with the clear goal to create datasets that can help organic reaction prediction. Governing committee around participants from industry (Google, Merck, Pfizer) and academia (MIT, Princeton)
XDL ²⁵		Developed to compile machine-readable experimental scripts that can be executed by robots
OECD Harmonised Templates	Harmonised Templates	data formats for information on chemicals and safety information
Materials Schema		effort to extend schema.org (a large effort that creates schema for structured data on the web) which would allow indexing by major search engines. This effort is inspired by the bioschemas initiative that has been developing types for the life sciences to increase the findability of data. Currently being developed (pre-alpha version) by NIST
Allotrope Data Format (ADF)		consortium of pharmaceutical industries and a larger partner network of instrument manufacturers with the aim to create a data infrastructure that captures the full lifetime of a sample in the lab. It has been working on an ontology (that is available under a permissive license), standard data schema/file and already has good coverage for many analytical techniques. Many developments (like access to the source code) require membership in the consortium
Unified Data Model (UDM)		format to store reaction data (schema, conditions, provenance, etc.) and biological testing. Has been maintained by Elsevier, who transferred the ownership to the Pistoia Alliance.



Supplementary Figure 1 | Non comprehensive schematic timeline of standards that have been proposed for chemical data. Inspired by Anderson et al.²⁶.

While there have been many efforts to create standards for data management in chemistry (some examples are described in Supplementary Tab. 1 and 2) there are only few (for example, the CIF) that are widely used. Motivated by this, there have been some initiatives that try to define standards for the field (see Supplementary Tab. 3). However, some of the most advanced efforts are being spearheaded by industry consortia, like the Allotrope Framework, which in some cases require membership to see the full documentation. While this can be a way to fund the efforts, it can also undermine one of the purposes for standardisation, i.e., to make data reusable by everyone.²⁷ We have to realise that one of the features of open source is that it can level the playing field, i.e., give less well-funded groups the same access to high-quality tools. Moreover, in the recent pandemic, we could witness how access to data can make a difference in response to an emergency.^{28;29}

From the computational materials science community we know that progress on open standards, even though slow, is possible. Starting 2016, the OPTIMADE³⁰ consortium, in which all major computational databases have been represented, agreed upon a specification with which the databases can “talk” to each other. Practically, this means that users can easily retrieve, combine, and compare data from all major databases.

Supplementary Table 2 | Examples of common characterisation techniques, file formats, and the conversion tools that have been developed by the cheminfo team (GitHub organisations cheminfo, cheminfo-js, cheminfo-py, mljs, image-js).

technique	file formats	conversion library
powder x-ray diffraction (PXRD)	Bruker brml, PowDLL xy	xrd
adsorption isotherms	Belsorp xls, DVS csv, Micrometrics csv and txt, IGA txt	isotherm-analysis
thermal gravimetric analysis (TGA)	Perkin Elmer csv and txt, TA Instruments txt	tga-spectrum
x-ray photoelectron spectroscopy (XPS)	VAMAS	vamas
nuclear magnetic resonance spectroscopy (NMR)	Bruker zip, JEOL jdf	brukerconverter, jeolconverter
liquid/gas chromatography (LC/GC)	NetCDF	netcdfjs
mass spectrometry (MS)	mzData, mzML, mzXML	mzData
images (transmission/scanning electron microscopy)	TIFF, png	tiff, fast-png

Supplementary Table 3 | Some initiatives that work on the standardisation, digitisation, and reusability of chemical data.

initiative	description
GOFair Chemistry Implementation Network	one of the ambitious goals is to create a management structure for standards through IUPAC, which has several ongoing projects concerning the developments of data standards ³¹
Research Data Alliance Chemistry Research Data interest group	aims to establish standards and ontologies and also proposes to involve instrument manufacturers in the discussions
RDA/CODATA Materials Data, Infrastructure & Interoperability interest group	aims to foster exchange of computational and experimental materials data via interoperability and shared standards
Materials Research Data Alliance	emerged from NSF's 2019 Summit on Big Data and Materials Cyberinfrastructure with the goal of creating a community around materials data sharing and reuse
National Research Data Infrastructure for Chemistry (NFDI4Chem)	German consortium with the aim to build an open and FAIR infrastructure for research data management in chemistry

Supplementary Table 4 | There is no lack of standards. This table is a non-comprehensive overview. The most common serialization language (serial.) is XML. XML is a web standard that is supported in most programming languages, is easily extendable, and is human-readable. It was originally developed as markup language and not as data storage format. In contrast to XML, JSON is relatively easy to parse, and it is also supported by most, if not all, programming languages (often having a direct mapping to standard datatypes such as dictionaries). Self-defining Text Archive and Retrieval (STAR) is the file structuring used in CIF. Not human-readable are HDF5 (Hierarchical Data Format), netCDF (which in the most recent implementation builds on top of HDF5), SQLite, and protocol buffers. Those file types have the advantage that they can represent data more compactly than the text-based format. The reason for that is, for example, the numerical accuracy of numbers. In string-based representation, it takes more space (to “print” all the digits) than in a native representation. One difference between the standards in this table is the extent to which the formats provide controlled vocabularies. Some only provide basic implementations like parameter labels with which any parameter can be stored, whereas others specify a tightly controlled vocabulary that only allows specific terms (blurring the line between format and schema). Only a few reuse and extend vocabularies that are more widely used, e.g., the web in general (schema.org). This overview clearly reflects that there will not be a one-size-fits-all format given that some applications need formats that are optimised for performance whereas others will require audits trails and digital signature features. But the table shows that many implementations are not that different, e.g., multiple HDF5 implementations of the mzML standard and that the community most likely does not need one additional (potentially backward incompatible) re-implementation of the mzML standard. What is interesting to observe is that most formats do not have a formalised mechanism for the proposal and discussion of changes to the standard. While some use public mailing groups and issue trackers, to our knowledge, only autoprotocol follows the schema of “enhancement proposals”, which is used in major software developments to propose and discuss new designs in a standardised, public-facing, and documented way. The inspection of the table also highlights that formats with buy-in from industry (e.g., ADF, UDM, AniML, Autoprotocol) found significantly wider adoption and are better maintained.

name	year	scopes	description	serial.
JCAMP-DX ³²	1988	general analytical chemistry	IUPAC recommended format. Many instruments can export this format, many databases such as the NIST web-book understand this format. Allows for some compression. Can be customised using private labels.	its own
ANDI ³³	1992	general analytical chemistry	The Analytical Data Interchange format is most used in its flavor for mass spectrometry, ANDI-MS more widely used. Designed by ASTM	netCDF
CML	1995	chemical data in general	Is supported by many common programs and has been extended, for example, for spectra (CMLSpect) ²⁴ . One core idea of designers of the CML was that the information should be validatable, for example, to automatically find if there negative atom counts	XML

NMRSTAR ³⁴	1996	(biological) NMR	archival format used by the Biological Nuclear Magnetic Resonance data Bank (BMRB), the international repository of biomolecular NMR data	STAR
matML ³⁵	1999	materials property data	Development initiated at NIST with the goal to develop a format for the interchange of materials information, but did not find wide usage	XML
SpectroML ³⁶	2002	UV-VIS	super-seeded by AnIML	XML
ThermoML ³⁷	2003	thermophysical and thermochemical property data	IUPAC standard for storage and exchange of experimental thermophysical and thermochemical property data. Data produced by the NIST Thermodynamics Research Center are provided in this format. IUPAC-based terminology is used as basis for data tagging	XML
GAML ³⁷	2003	general analytical chemistry	The Generalized Analytical Markup Language attempted to make it easier to store multi-detector data (compared to JCAMP and ANDI). Explicitly avoids mapping vendor specific metadata to a common dictionary	XML
mzXML ³⁸	2004	mass spectrometry	strict schema with enumerated attributes, officially deprecated in favour of mzML	XML
AnIML ¹⁷	2004	general analytical chemistry	The Analytical Information Markup Language aims to document workflows. It allows to embed digital signatures and audit trails, aims to be flexible enough for novel spectroscopic techniques. Found already support by some manufacturers like Agilent or ELNs such as LabWare. BSSN software (now owned by Merck) implemented converters for more than 150 instruments ¹⁸	XML
mzData ³⁹	2006	mass spectrometry	Designed to be flexible via extendable controlled vocabulary	XML
mzML ⁴⁰	2008	mass spectrometry	Aims to combine the best elements of mzXML and mzData, metadata is accurately and unambiguously annotated using the PSI-MS controlled vocabulary	XML
mz5 ⁴¹	2012	mass spectrometry	Uses HDF5 in an attempt to address performance issues with mzML. Manual mapping to the field labels of mzML	HDF5
ADF	2013	general analytical chemistry	The Allotrope Data Format features an ontology based on the widely used Basic Formal Ontology. Data is stored in RDF graph, can be programatically validated. The format also supports audit trails. Additional files can be stored in virtual file system. Full analytical life-cycle can be stored in one file. Wide support from instrument manufactures. Maintained by the Allotrope Foundation, which was founded by pharmaceutical companies, and which uses the membership fee paid by industrial members to contract external partners to develop and maintain the format	HDF5
mzDB ⁴²	2015	mass spectrometry	Attempts to optimise for high-throughput data processing and storing	SQLite
autoprotocol ⁴³	2016	life science protocols	Autoprotocol is directly mappable to hardware commands for robotic automation. Notably, it has a mechanism for the proposal of changes to the standard (Autoprotocol Standard Changes). It is maintained by Strateos	JSON
PIF	2017	information about physical systems	The Physical Information File was designed to be able to describe a broad array of data "from parts in a car down to a single monomer in a polymer matrix", hence can have a nesting of subsystems. It is maintained by Citrine Informatics	JSON

UDM	2017	experimental information about compound synthesis and testing	The Unified Data Model is a format to store reaction data (schema, conditions, provenance, etc.) and biological testing. Has been maintained by Elsevier, who transferred the ownership to the Pistoia Alliance.	XML
NMReDATA ⁴⁴	2018	NMR	compared to nmrML and NMRSTAR tries to not be exhaustive but limit itself to core set of parameters, carefully include structure assignment. Maintained by the NMReDATA initiative	SDF
nmrML ⁴⁵	2018	NMR	one goal was to follow the design of mzML, the motivation was to address the different "dialects" of JCAMP-DX implementations	XML
GEMD	2019	links materials, the processes that produce them, and the measurements that characterize them. It resolves some problems that were present in the data model underlying the PIF	JSON	
ord-schema	2019	organic reactions	The developers built this schema based on a survey they conducted in late 2019/early 2020 with the clear goal to create datasets that can help organic reaction prediction. The consortium is lead by a Governing Committee with members from industry (Google, Merck, Pfizer) and academia	protocol buffer
XDL	2020	organic reaction protocols	Developed to compile machine-readable experimental scripts that can be executed by robots such as the "chemputer". ²⁵	XML
AIF ⁴⁶	2021	gas de/adsorption isotherms	The authors of the Adsorption Information File provide tools convert proprietary files to AIF	STAR
mzMLb ⁴⁷	2021	mass spectrometry	Uses HDF5 in an attempt to address performance issues with mzML, preserves the mzML structure. Attempts to improve over mz5 by preserving the link to also future mzML versions	HDF5

Glossary

access token	are random character strings that can, similar to passwords, be used to grant the rights for certain operations. Usually, access tokens can be minted for a limited scope, e.g., for only one sample and only specific rights (e.g., read rights).
API	an Application Programming Interface (API) describes how one can interact with a program. It describes what requests can be made and in which form they have to be made. In contrast to a user interface it is not intended for direct use via a “frontend” but rather to offer programmatic access.
backend	the piece of software that accesses the data. One standard model in software development is the client-server model, where the backend is the part of the code that runs on the server.
client	the archetypal example of a client is a web browser. A client can be use to access a service offered by a server.
data dictionary	a detailed definition (format, meaning, relationship, usage) and documentation about each data element (in a data model).
data lake	central place for collecting raw data from different sources. In the context of ELNs this can refer to system that simply stores all raw, unparsed data that might find potential use at some later point in time.

data mesh	data meshes are, in contrast to centralised data lakes or data warehouse, a decentralized data architecture. Each domain will offer data as a product. ⁴⁸ This can be thought of as the data equivalent of microservices that found broad use for scalable software products..
data model	abstract description of the data structure (the elements and their relationships). It is more higher level than schema, one example is the relational model, another is a graph model.
data warehouse	central place for structured data integrated from different sources. In the context of an ELN, the term data warehouse might be used to describe the system that is used to integrate data captured from different laboratories, on different instruments.
deployment	deployment refers to the activities that are needed to make a piece of software available to users.
disruptive innovation	a disruptive innovation is an innovation that creates a new market. One example for this are smartphones.
DMP	data management plans describe how data are handled during and after a research project.
DOI	a digital object identifier (DOI) is a <i>unique and persistent</i> identifier that is created by some central agency that keeps track of some metadata and commits to resolving to it indefinitely.
Dublin Core	the Dublin Core Metadata Element Set is a set of 15 elements like "Contributor", "Date", "Format" that should be used for describing data resources. A similar set of metadata is the DataCite Metadata Schema.

ELN	in this perspective we see an electronic lab notebook (ELN) not only as a place where experimental chemists take notes but as a hub for chemical research. Note taking is an important functionality in this hub, but visualising, interacting, and sharing data are at least equally important.
FAIR	the FAIR principles are guidelines on how the findability, accessibility, interoperability, and reusability of data can be improved. ⁴⁹ Importantly, FAIR does not necessarily mean open.
five star data	Tim Berners-Lee suggests that the highest-quality open and linked data, i.e. 5-star data, ⁵⁰ must (1) be available on the web under an open license (e.g., creative commons), (2) be structured (e.g., table instead of the image of a table), (3) be non-proprietary (e.g., using standard text files instead of binary formats), (4) use uniform resource identifiers (URIs) (e.g., using an URL such that others can share and link to this piece of data), (5) link to other data. In contrast to FAIR, linked open data (5-star data) must be open.
frontend	the piece of software that displays the data, it is the part of a website the users interact with. In the client-server model this is the piece of software that runs on the client (e.g., the web browser).
GitHub	github.com is where most of the open source code “lives” and is collaboratively built. It is integrated with the version control tool Git but also give the option to make copies of other code bases (“fork”/“clone”) or to contribute code to some projects using “pull requests”. Bugs and new features are usually discussed in “issues”.

JavaScript	JavaScript is a programming language that is at the core of most of the techniques of the world wide web and can be understood by most web browsers.
Kaggle	is the largest community for data science and most known for its competitions in which participants compete to create the best model for a task posed by the host of the competition.
LIMS	laboratory infrastructure management systems allow to track and manage samples. Often they allow to define locations at which different material or equipment is stored and can then track the moves and uses of the materials.
link rot	link rot describes the phenomenon that links tend to cease to resolve to the content they originally pointed to.
markup	a markup language allows annotating the text to decouple the structure from the content. One example is $\text{\LaTeX}$ where the <code>\section</code> command can be used to define the concept of a section. In the context of data, markup languages such as Extensible Markup Language (XML) are typically used. The archetypal example for chemistry is CML.
metadata	metadata is data that describes attributes of other data and in this way provides context and makes data discoverable. While this might sound abstract there are many use cases in which metadata can be helpful. For example, date and time should be featured in most metadata records. If we have access to them we can use them to filter our data, but also correlate them with other information (such as the weather reports) to understand potential outliers in our data. A useful overview of metadata schemes is https://rdamsc.bath.ac.uk/ .

ontology

an ontology describes and clarifies the meaning, conceptualisation, and relation of terms in a formal language.⁵¹ That is, they go beyond the a simple vocabulary or thesaurus by also providing a model for the information. Ontologies are often written in the web ontology language (OWL). For example in the ChEBI ontology⁵² “metal atom” is defined as “An atom of an element that exhibits typical metallic properties, being typically shiny, with high electrical and thermal conductivity” and the relation to other terms is formally described (i.e., on could make mathematical inferences, e.g., on the consistency). In the context of chemical data and ELNs ontologies are important to make sure that it is clear and well defined what certain items in a database mean (for example, the concept “bond order” depends on the definition—for this reason a “bond order” entry in a database would link to a ontology with the definition). Importantly, ontologies are relevant to enable semantic interoperability, i.e., enabling the decoding of the meaning of data. Pachl et al.⁵³ give an excellent overview over ontologies that have been developed for chemistry (and find that many projects remain unfinished). One example that provides a controlled vocabulary that can be programmatically accessed using an application programming interface (API) is the IUPAC gold book. However, since a vocabulary is not a full ontology, there are ongoing efforts to also model the relationship between terms to arrive at a full chemical ontology.⁵⁴

open notebook science	popularized by Jean-Claude Bradley, ⁵⁵ open notebook science refers to making <i>all</i> primary research outputs openly available as it they are recorded. The goal is to minimise the amount of “dark data”, e.g., “unsuccessful” experiments that would never get published, and “insider information” that one might not report in methods sections in papers. openlabnotebooks.org collects some (mostly life-science) efforts.
open source	open source software is software that is openly available for modification and reuse. Development of open source software is often decentralised, and happens in open collaboration.
<i>Organic Syntheses</i>	<i>Organic Syntheses</i> is a journal that publishes procedures for the synthesis of organic compounds. The key distinction is that the syntheses are replicated in a lab of one of the members of the editorial board.
proprietary	proprietary files contain data that are stored in an encoding scheme defined by an organisation. The difficulty with proprietary files stems from the fact that the schema is secret, that is, it is often not trivial to decode and interpret the files. For this reason, data can only be five star data if it is shared in a non-proprietary format.
provenance	data provenance is a form of metadata that describes the origin of data.
RDF	Resource Description Framework (RDF) triple-stores break down information into triples of “subject”, “verb”, “object” that form a labeled graph which machines can use to explore information (and understand how information is connected). An example for a triple can be “substance” (subject) “inhibits” (verb) “protein” (object) and the same element can also be used in “publication” (subject) “describes” (verb) “protein”.

relational	relational databases store table in tables, comparable to spreadsheets with well-defined columns. The word “relational” refers to the fact that the rows within a table are related to each other according to some concept such as “chemicals”.
repository	is a location where software (in the case of software repositories, e.g., on GitHub) or data are stored/archived (e.g., Zenodo). www.re3data.org/ is a useful resource to find suitable repositories.
REST	(Representational State Transfer (REST)). One key design principle of RESTful interfaces is that they are stateless. Practically speaking this mean that the clients will send <i>all information</i> that is needed to process a request to a <i>server</i> . That is, the client, and not the server, stores all the information about the client state. Importantly, the REST scheme also defines a uniform structure for the APIs.
schema	a schema describes the structure of the data in a formal language. JSON-LD and XML are common formats for schemas. Examples in the context of the sciences are XML schema of the Chemical Markup Language (CML) ²² or the Analytical Information Markup Language (AnIML) ¹⁷ .
semantic web	the semantic web ⁵⁶ is an extension of the world wide web with metadata and data descriptors that makes data machine-readable and, more important, machine-interpretable such that intelligent digital agents can perform actions based on the data.
server	is the computer (program) that provides some functionality for the clients. A typical example is a file server that provides files to its client (as it is for example done in companies). One typical server that is used in ELNs is the database server that serves the database to the clients.

trusted timestamps	trusted timestamps proof that certain information existed at a certain point in time. Importantly, they are secure in the sense that also the owner of the document cannot change the timestamp. They can be secured by techniques like the blockchain ⁵⁷ or a time-stamping authority.
URI	Uniform Resource Identifiers (URIs, as defined in RFC 3986) are sequences of characters that identify a physical (e.g., sample) or digital (e.g., web page) resource. One special example are Uniform Resource Locators (URLs). Internationalized Resource Identifiers (IRIs) are generalisations of URIs to unicode glyphs. Note that there are initiatives, such as the Resource Identification Portal, that provide a centralised location for URIs for research resources such as antibodies or cell lines.
UUID	a universally unique identifier (UUID) is a 128-bit number that is, with high probability, globally unique. That means that this number is so long that there is a practically zero chance that someone else uses the same number to identify something else.
web browser	in an abstract sense, a web browser is a piece of software that can be used to retrieve information from the world wide web. Practically speaking it can “speak” certain programming languages like JavaScript and HTML. This means that a program that is written in JavaScript can be executed by all major browsers on any platform or any device. This led some people to say that a web browser is a “super app” ⁵⁸ that could potentially replace almost all other apps on a computer. This is important in the context of ELNs as this might allow us to create systems that are easier to scale (one does not have develop a separate app for every operating system).

web service

a web service is a program that makes itself accessible via the world wide web.

References

1. Gabriel, R. The rise of “worse is better”. *Lisp: Good News, Bad News, How to Win Big* **1991**, 2.
2. Stonebraker, M. Tamr White Paper – The Seven Tenets of Scalable Data Unification. **2019**, 13.
3. Oleksik, G.; Milic-Frayling, N.; Jones, R. Study of electronic lab notebook design and practices that emerged in a collaborative scientific environment. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14. 2014.
4. Raymond, E. *The cathedral & the bazaar : musings on Linux and open source by an accidental revolutionary*; O'Reilly: Beijing, 1999.
5. Turner, J. Open Source Has a Funding Problem. <https://stackoverflow.blog/2021/01/07/open-source-has-a-funding-problem/>, 2021.
6. Stannard, A. How to Build Sustainable Open Source Software Projects. <http://www.aaronstannard.com/sustainable-open-source-software/>, 2020.
7. Gowers, T.; Nielsen, M. Massively collaborative mathematics. *Nature* **2009**, 461, 879–881.
8. Woelfle, M.; Oliaro, P.; Todd, M. H. Open science is a research accelerator. *Nat. Chem.* **2011**, 3, 745–748.
9. Epstein, D. *Range : why generalists triumph in a specialized world*; Riverhead Books: New York, 2019.
10. Gowers, T. Questions of Procedure. 2009.
11. Inc, K. Kaggle: Your Home for Data Science. <https://www.kaggle.com/>, 2021.
12. Hanwell, M. D.; de Jong, W. A.; Harris, C. J. Open chemistry: RESTful web APIs, JSON, NWChem and the modern web application. *J. Cheminf.* **2017**, 9.
13. A. Badiola, K. et al. Experiences with a Researcher-Centric ELN. *Chem. Sci.* **2015**, 6, 1614–1629.

14. Jablonka, K. M.; Moosavi, S. M.; Asgari, M.; Ireland, C.; Patiny, L.; Smit, B. A data-driven perspective on the colours of metal–organic frameworks. *Chem. Sci.* **2021**,
15. Brooks, B. J.; Thorn, A. L.; Smith, M.; Matthews, P.; Chen, S.; O'Steen, B.; Adams, S. E.; Townsend, J. A.; Murray-Rust, P. Ami - The chemist's amanuensis. *J. Cheminf.* **2011**, 3.
16. Knight, N. J.; Kanza, S.; Cruickshank, D.; Brocklesby, W. S.; Frey, J. G. Talk2Lab: The Smart Lab of the Future. *IEEE Internet of Things Journal* **2020**, 7, 8631–8640.
17. Schäfer, B. A.; Poetz, D.; Kramer, G. W. Documenting laboratory workflows using the Analytical Information Markup Language. *JALA* **2004**, 9, 375–381.
18. Schäfer, B. Data Exchange in the Laboratory of the Future. *Wiley Analytical Science* **2018**,
19. Miles, B.; Lee, P. L. Achieving Reproducibility and Closed-Loop Automation in Biological Experimentation with an IoT-Enabled Lab of the Future. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* **2018**, 23, 432–439.
20. Michel, K.; Meredig, B. Beyond bulk single crystals: A data format for all materials structure–property–processing relationships. *MRS Bulletin* **2016**, 41, 617–623.
21. Frenkel, M.; Diky, V.; Chirico, R. D.; Goldberg, R. N.; Heerklotz, H.; Ladbury, J. E.; Remeta, D. P.; Dymond, J. H.; Goodwin, A. R.; Marsh, K. N.; Wakeham, W. A.; Stein, S. E.; Brown, P. L.; Koenigsberger, E.; Williams, P. A. ThermoML: an XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermo-physical and Thermochemical Property Data. 5. Speciation and Complex Equilibria. *Journal of Chemical & Engineering Data* **2011**, 56, 307–316.
22. Murray-Rust, P.; Rzepa, H. S. CML: Evolution and design. *J. Cheminf.* **2011**, 3.
23. Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of chemical markup language (CML) as a system for handling complex chemical content. *New J. Chem.* **2001**, 25, 618–634.

24. Kuhn, S.; Helmus, T.; Lancashire, R. J.; Murray-Rust, P.; Rzepa, H. S.; Steinbeck, C.; Willighagen, E. L. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. *J. Chem. Inf. Model.* **2007**, *47*, 2015–2034.
25. Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **2020**, *370*, 101–108.
26. Anderson, A.; McGibbon, G.; Paramonov, A.; Bhal, S. *Looking Beyond Analytical Data Standardization - the Fourth Paradigm*; p 8.
27. Davies, A. N.; Lancashire, R. How Standard Are Your Standards? *Spectrosc. Eur.* **2017**, *29*, 3.
28. Bruno, I.; Frey, J. G. Connecting Chemistry with Global Challenges through Data Standards. *Chem. Int.* **2017**, *39*, 5–8.
29. Aitsi-Selmi, A.; Blanchard, K.; Murray, V. Ensuring science is useful, usable and used in global disaster risk reduction and sustainable development: a view through the Sendai framework lens. *Palgrave Communications* **2016**, *2*.
30. Andersen, C. et al. The OPTIMADE Specification. 2020; <https://doi.org/10.5281/zenodo.4195051>.
31. McEwen, L. R. Towards a Digital IUPAC. *Chem. Int.* **2020**, *42*, 15–17.
32. Grasselli, J. G. Jcamp-Dx, a Standard Format for Exchange of Infrared Spectra in Computer Readable Form. 2016.
33. Matthews, L.; Miller, T. ASTM Protocols for Analytical Data Interchange. *JALA: Journal of the Association for Laboratory Automation* **2000**, *5*, 60–61.
34. Ulrich, E. L.; Argentar, D.; Klimowicz, A.; Westler, W. M.; Markley, J. L. STAR/CIF macromolecular NMR data dictionaries and data file formats. *Acta Crystallographica Section A Foundations of Crystallography* **1996**, *52*, C577–C577.
35. Varde, A. S.; Begley, E. F.; Fahrenholz-Mann, S. MatML: XML for Information Exchange with Materials Property Data. Proceedings of the 4th International Workshop on Data Mining Standards, Services and Platforms. New York, NY, USA, 2006; p 47–54.

36. Rühl, M. A.; Schaefer, R.; Kramer, G. W. Spectro ML-A Markup Language for Molecular Spectrometry Data. *JALA: Journal of the Association for Laboratory Automation* **2001**, 6, 76–82.
37. Frenkel, M.; Chirico, R. D.; Diky, V. V.; Dong, Q.; Frenkel, S.; Francois, P. R.; Embry, D. L.; Teague, T. L.; Marsh, K. N.; Wilhoit, R. C. ThermoMLAn XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data. *Journal of Chemical & Engineering Data* **2003**, 48, 2–13.
38. Pedrioli, P. G. A. et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology* **2004**, 22, 1459–1466.
39. Orchard, S.; Hermjakob, H.; Taylor, C.; Binz, P.-A.; Hoogland, C.; Julian, R.; Garavelli, J. S.; Aebersold, R.; Apweiler, R. Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4–6, 2005. *PROTEOMICS* **2006**, 6, 738–741.
40. Deutsch, E. mzML: A single, unifying data format for mass spectrometer output. *PROTEOMICS* **2008**, 8, 2776–2777.
41. Wilhelm, M.; Kirchner, M.; Steen, J. A.; Steen, H. mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets. *Molecular & Cellular Proteomics* **2012**, 11, O111.011379.
42. Bouyssié, D.; Dubois, M.; Nasso, S.; de Peredo, A. G.; Burlet-Schiltz, O.; Aebersold, R.; Monsarrat, B. mzDB: A File Format Using Multiple Indexing Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets. *Molecular & Cellular Proteomics* **2015**, 14, 771–781.
43. Bates, M.; Berliner, A. J.; Lachoff, J.; Jaschke, P. R.; Groban, E. S. Wet Lab Accelerator: A Web-Based Application Democratizing Laboratory Automation for Synthetic Biology. *ACS Synthetic Biology* **2016**, 6, 167–171.
44. Pupier, M. et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magnetic Resonance in Chemistry* **2018**, 56, 703–715.
45. Schober, D. et al. nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. *Analytical Chemistry* **2017**, 90, 649–656.

46. Evans, J. D.; Bon, V.; Senkovska, I.; Kaskel, S. A Universal Standard Archive File for Adsorption Data. *Langmuir* **2021**, *37*, 4222–4226.
47. Bhamber, R. S.; Jankevics, A.; Deutsch, E. W.; Jones, A. R.; Dowsey, A. W. mzMLb: A Future-Proof Raw Mass Spectrometry Data Format Based on Standards-Compliant mzML and Optimized for Speed and Storage Requirements. *Journal of Proteome Research* **2020**, *20*, 172–183.
48. Dehghani, Z. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. <https://martinfowler.com/articles/data-monolith-to-mesh.html>, 2019.
49. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*.
50. Hyland, B.; Atemez, G.; Pendleton, M.; Srivastava, B. Linked Data Glossary. <https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html> #x5-star-linked-open-data, 2013; accessed Jan. 8, 2021.
51. Uschold, M. Ontology and database schema: What's the difference? *Appl. Onto.* **2015**, *10*, 243–258.
52. Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2007**, *36*, D344–D350.
53. Pahl, C.; Frank, N.; Breitbart, J.; Bräse, S. Overview of Chemical Ontologies. *arXiv:2002.03842 [cs]* **2020**,
54. Chalk, S. The IUPAC Gold Book: Toward a Chemical Ontology for the Next 100 Years of Chemistry. <https://iupac.org/100/stories/the-iupac-gold-book/>, 2019.
55. Bradley, J.-C. Open Notebook Science Using Blogs and Wikis. *Nat. Preced.* **2007**,
56. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Amer.* **2001**, *284*, 34–43.
57. Gipp, B.; Meuschke, N.; Gernandt, A. Decentralized Trusted Timestamping using the Crypto Currency Bitcoin. Proceedings of the iConference 2015. Newport Beach, CA, USA, 2015.

58. Lex Fridman, Brendan Eich: JavaScript, Firefox, Mozilla, and Brave | Lex Fridman Podcast
#160. 2021; <https://www.youtube.com/watch?v=krB0enBeSiE>.