Characterization Of The Willard L. Eccles Observatory For Optical Astronomy

Article · May 2011

CITATIONS
0

READS
10

4 authors, including:

Dennis Della Corte
Brigham Young University - Provo Main Campus
20 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Wayne Springer
University of Utah
30 PUBLICATIONS **470** CITATIONS
SEE PROFILE

Drug Discovery Today • Volume xxx, Number xx • xxxx 2021 REVIEWS

ELSEVIER

Self-reporting data assets and their representation in the pharmaceutical industry

Dennis Della Corte [a,b,*], Wolfgang Colsman [b], Heiko Fessenmayr [c], Alexandre Sawczuk da Silva [b], Dana E. Vanderwall [d]

[a] Department of Physics and Astronomy, Brigham Young University, UT, USA
[b] ZONTAL, Inc, Melbourne, FL, USA
[c] R&D Software Informatics, Agilent, USA
[d] Research & Early Development IT, Bristol Myers Squibb, USA

Standardizing data is crucial for preserving and exchanging scientific information. In particular, recording the context in which data were created ensures that information remains findable, accessible, interoperable, and reusable. Here, we introduce the concept of self-reporting data assets (SRDAs), which preserve data and contextual information. SRDAs are an abstract concept, which requires a suitable data format for implementation. Four promising data formats or languages are popularly used to represent data in pharma: JCAMP-DX, JSON, AnIML, and, more recently, the Allotrope Data Format (ADF). Here, we evaluate these four options in common use cases within the pharmaceutical industry using multiple criteria. The evaluation shows that ADF is the most suitable format for the implementation of SRDAs.

Keywords: Data format; ADF; Scientific data; JSON; JCAMP-DX; Pharmaceutical data; AnIML; FAIR data; Ontology; Data model

Introduction

Exchanging technical information is a crucial part of data-intensive scientific domains, such as the pharmaceutical research community.[1] Recent reports highlight the importance of standardized data for the successful digital transformation of the pharmaceutical industry.[2,3] Despite the promised benefits for pharmaceutical companies, many have not used standardized data formats at scale to date because of the significant change that is involved because of the quantity and complexity of the data involved; instead, they frequently opt to using proprietary systems or even Excel spreadsheets.[4] Clearly, this lack of standardized data significantly hinders the research process because data maintained under such conditions are less likely to adhere to FAIR principles, which ensure that the data are findable (F), accessible (A), interoperable (I), and reusable (R).[5,6] Without observing these guidelines, it becomes difficult to share, process, and analyze data in a meaningful way across laboratories and organizations.

Here, we address the issue of standardizing data by proposing the concept of SRDAs, which fulfill a set of content and technical requirements that enable important business benefits in the pharmaceutical industry. SRDAs are an efficient representation of data and descriptive information that enables full interpretation of data assets without external dependencies. For instance, a SRDA that records the result of a particular laboratory experiment will also contain information about experimental settings, equipment, and the overall project. By maintaining this contextual information, SRDAs support the FAIR principles of findability, interoperability, and reuse; ensure auditability by tracking how data are collected and processed;[7] and establish data integ-

* *Corresponding author at: Department of Physics and Astronomy, Brigham Young University, UT, USA.* Della Corte, D. (Dennis.dellacorte@byu.edu)

https://doi.org/10.1016/j.drudis.2021.07.019

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

rity through consistent handling and long-term preservation.[8] This approach can capture data with different levels of granularity, catering for different payloads and semantic models.

SRDAs are introduced here at the conceptual level; thus, it is necessary to find a suitable data format or language to implement them. Several specific formats and languages for the reuse and long-term preservation of scientific data in the pharmaceutical domain have been proposed, with four distinct strategies for structuring information: JCAMP-DX;[9] a flavor of JSON (JSON is a technology widely used in the application landscape. Although JSON is a file format and not a data standard, we include it in the comparison as a proxy to represent various solutions built on this technology);[10]) AnIML (an XML-based format proposed in 2004[11,12] and ADF (established in 2017,[13] and based on HDF5,[14] a scientific data format).[15] The first three annotate scientific data using labels (in the case of JCAMP-DX), keys (in the case of JSON), or tags (in the case of XML), which provide metadata describing the experimental results. ADF is structured in three layers: recording metadata, long-term readable data, and the original information in its raw form. All these models are focused on capturing semantic information, which ties the data to the context in which they were created; in the case of the pharmaceutical lab, this is the equipment and processes within the laboratory.[16,17] Having this contextual information enables a data set to be understood by a broader audience that is not an expert in the workflow that created it, such as a data scientist who wants to reuse chromatography data sets. Two other formats were considered but ultimately excluded from our analysis: mzML, because its development stopped in 2017, and JSON-LD, because the authors found limited implementation reports from scientific laboratories. Other legacy formats with limited relevance include CDF and netCDF.[18,19]

Here, we introduce SRDAs and analyze which of the four data formats can best support the implementation of them. To achieve this goal, requirements for SRDAs are extracted from common use cases in the pharmaceutical industry. Then, a fit-gap analysis for each requirement is conducted to determine the most suitable format to fulfil it. We then discuss typical use cases in the pharmaceutical industry, focusing on the requirements they pose for SRDAs, introduce the data formats discussed in this analysis; perform a fit-gap analysis for the chosen requirements; and then summarize and discuss the analysis.

Evaluation criteria from pharmaceutical industry-use cases

The requirements that an SRDA needs to fulfill can be derived from the broad number of pharmaceutical-use cases in which SRDAs are deployed. One example is the electronic archiving of experimental raw data, together with the reports created to interpret them. Multiple top pharma companies have been working to answer the question of how to effectively represent data, aiming to both abide by regulatory demands and preserve the information in a way that facilitates future reuse.[2] The use cases discussed herein capture the complexity of this undertaking and reveal several specific points of focus.

Open archival information system

One frequent use case in regulated industries is the need to preserve data longer term for compliance reasons. The leading industry standard Open Archival Information System (OAIS) is used for long-term archival by multiple major organizations.[20] The OAIS framework introduces the concept of an information package, which creates a logical link between information and raw data. In this context, the requirements for SRDAs are that the raw data representation is persistent in the long term, the associated metadata are of high quality (i.e., are easy to find and provide meaningful context), and an audit trail exists to ensure the continual tracking of changes to the SRDA.

Data-centric IT landscape

Another common use case concerns the creation of a data-centric IT landscape within a large enterprise (such as Novartis). To effectively connect and exchange information, SRDAs need to fulfill several requirements: data representations should be created in a format that is easy to generate and exchange among systems; this format also needs to be extensible, so that it can collect and integrate other descriptive information as required. As an example, for the second requirement, a data entry from a Laboratory Information Management System (LIMS) might need to be made available in a data science dashboard; however, the LIMS does not contain a definition of the associated project. In this situation, a workflow needs to be able to extract additional project information from a master data system and integrate it into the LIMS data export. Native vendor formats are often difficult to enrich in such a fashion. To solve this, an SRDA needs to standardize information across a diverse set of data sources and be able to be rapidly enriched with additional descriptive information. In many cases, this process must also be governable through GxP validated workflows.[21]

Instrument data integration

Integration of raw data from laboratory instruments, such as ELISA, and downstream data analysis tools, such as dashboards, is a common problem in smaller biotech companies as well as large pharma. However, without the IT infrastructure of large pharmaceutical organizations, independent small companies use SRDAs to enable project-wide analysis. The main requirements in this scenario are that homegrown systems can be quickly updated to produce data according to SRDA standards. SRDAs also need to be well structured, enforcing data quality that allows for a successful merger and acquisition of the biotech in the future.

Harmonization of metadata and raw data

The harmonization of raw data from mainstream laboratory systems into SRDAs (for data processing and archiving) is a typical use case for large pharmaceutical companies. Some of the requirements are to build SRDA representations for nuclear magnetic resonance (NMR) data, build SRDA representations for mass spectrometry data, and build SRDA representations for liquid chromatography (LC)-UV data. To accurately represent the highly complex machines producing these data, the requirement to accurately describe experimental conditions and raw data measurements also applies. Additionally, it is important to

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

ensure that the data models used are adhered to across different companies, ensuring that an SRDA can be truly self-reporting when opened in another context.

Collaboration between entities

A crucial use case in the industry is the exchange of information between contract research organizations (CROs) and pharmaceutical companies. Clearly, the requirement that SRDAs effectively communicate experimental details and measurements is vital, and the current use of PDF and EXCEL documents does not fully support this. As a practical implementation example for this requirement, one top pharma company is currently working on SRDA implementations that can solve this gap through a vendor-independent processing pipeline. Additionally, this company is developing data/metadata-extraction solutions to power visualizations. A collection of SRDAs is also expected to be useful in this context as a training set for the development of machine-learning models.

Available data container, languages, and standards

Preserving and exchanging scientific information is essential for research to flourish,[22] but this can only be achieved by preserving data using appropriate formats. Over time, different formats have been developed,[23] such as HDF5.[14] These were designed to provide methodical and thorough ways of structuring and maintaining records. However, in many cases, they ultimately impose limitations on the reuse, long-term preservation, and management of data, impacting research progress and the development of products. This is a crucial issue in the pharmaceutical industry, in which data must be preserved and made easily accessible for

TABLE 1

Example implementations in three standards or languages.

Example of a core portion of a JCAMP-DX file[9]	Example of a JSON file, in the format developed for pharmaceutical data[10]	Example of an AnIML file[27]

##TITLE= Epichlorohydrin vapor
##JCAMP-DX= 4.24
##DATA TYPE= INFRARED SPECTRUM
##ORIGIN= Sadtler Research Laboratories
##OWNER= EPA/Public Domain
$$ More optional lines can be included here
##XUNITS= 1/CM
##YUNITS= ABSORBANCE
##XFACTOR= 1.0
##YFACTOR= 0.001
##FIRSTX= 450.
##LASTX= 4000.
##NPOINTS= 1842
##FIRSTY= 0.058
##XYDATA= (X++(Y..Y))
450 58 44 34 39 26 24 22 21
21 19 16 15 15 17 16 etc.
3998 16 15 14
##END=
	{
	"@idsType": "cell-counter",
	"@idsVersion": "v1.0.0",
	"@idsNamespace": "common",
	"system": {"serial_number": "serial_number"},
	"run": {"id": "413befdd-c7e2-4edd-9e9b-06cf1cb0283f"},
	"time": {"measurement": "2015-09-24T03:47:13.0Z"},
	"sample": {
	"id": "unknown-10",
	"batch": {"id": "batch-number"}
	},
	"method": {
	"instrument": {"cell_type": "CHO","dilution_factor": 1}
	},
	"user": {"name": "operator-1"},
	"result": {
	"cell": {
	"viability": {"value": 0.1,"unit": "Percent"},
	"diameter": {
	"average": {
	"live": {"value": 21.07,"unit": "Micrometer"}
	}
	},
	"count": {
	"total": {"value": 1207, "unit": "Cell"},
	"viable": {"value": 1, "unit": "Cell"}
	},
	"density": {
	"total": {"value": 102.24, "unit": "MillionCellsPerMilliliter"},
	"viable": {"value": 0.1, "unit": "MillionCellsPerMilliliter"}
	}
	}
	}
	}
		<AnIML xmlns="urn:org:astm:animl:schema:core:draft:0.37"
		xmlns:xsi="http://www.w3.org/2001/XMLSChema-instance"
		xsi:schemaLocation="urn:org:astm:animl:schema:core:draft:0.37
		http://animl.cvs.sourceforge.net/*checkout*/animl/schema/animl-core.xsd?revision=1.72"
		version="0.37">
		
		<!-- SampleSet is defined in animl-core.xsd -->
		<SampleSet>
		<Sample name="Pond spike 1.5 ppm phos" sampleID="sample0001" id="ID000001">
		<!-- You can describe one or more samples here -->
		</SampleSet>
		
		<!-- ExperimentStepSet is defined in animl-core.xsd -->
		<ExperimentStepSet id="ID000002">
		<ExperimentStep name="UV/Vis Analysis" experimentStepID="STEP0001">
		<Infrastructure>
		<Method>
		<Result id="RESULT0001" name="UV/Vis Spectrum">
		</ExperimentStep>
		<!-- You can describe one or more samples here -->
		<ExperimentStepSet>
		
		<!--AuditTrail is defined in animl-core.xsd -->
		<AuditTrailEntrySet>
		<AuditTrailEntry>
		<!-- Over time the audit trail will grow when/if you change the file -->
		</AuditTrailEntrySet>
		
		<!-- SignatureSet is defined in the xmldsig-core-schema.xsd via animl-core.xsd -->
		<SignatureSet>
		<Signature>
		<!-- You can sign all or part of a document -->
		</SignatureSet>
		</AnIML>

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

research and regulatory purposes.[1] In this context, four data formats commonly used in this domain are further explored herein.

JCAMP-DX is a standard format developed during the late 1980s for exchanging data produced by spectrometers.[9] JCAMP-DX files are structured in blocks, which contain labeled records with spectral data. These blocks contain an associated scope, which captures the parameters used when producing these data. More extensive notes, which can be used to describe the equipment and observation method in more detail, can also be added. An example of a JCAMP-DX file is shown in Table 1. This format was designed to allow for spectral data to be consistently represented and exchanged, regardless of the make and model of the instrument that produced them.[9] However, it was not designed to satisfy the requirement of rapid searching over large amounts of data. Proprietary extensions of JCAMP-DX have been made available by different vendors, but those are not publicly documented.

JSON is one of the languages that has been proposed to tackle the complexity of scientific, in particular, pharmaceutical, data.[10] In this context JSON syntax is used to record the different data attributes.[24] An example of this type of JSON is provided in Table 1. In this paradigm, data are organized using key-value pairs, in which the value captures the data and the key contextually describes them. To maintain consistency across different semantic terms, this format uses JSON schemas,[10] which constrain which terms can be used for describing data. JSON can capture complex data, allowing for both single and multidimensional arrays.

AnIML is another data format that has gained traction in the pharmaceutical domain.[12] The acronym stands for Analytical Information Markup Language, indicating that the format is built upon XML.[25] An example of an AnIML file is provided in Table 1. AnIML is organized in two different layers. The first is the core layer, which holds data on results, samples, and audit trails. The second is the technique definition layer, which provides semantic metadata for the information in the core layer. This provides additional context, including experimental settings and data connections. In practice, the information of AnIML is controlled through the use of XML schemas, which dictate how data are organized.[26,27]

A precompetitive consortium of Pharma companies and vendors, the Allotrope Foundation, proposed the ADF to preserve scientific data, with a structure that is conducive to the encoding of pharmaceutical information based on HDF5.[13–15] The structure of an ADF file is illustrated in Fig. 1. ADF is organized around four different components. The first is data description, which contains metadata that preserve information regarding the provenance and type of data and is based on the resource description framework (RDF).[28] This description captures the context in which data were produced, including information about instruments, processes, samples, and results. The second is the data cube, which preserves analytical data that are represented either as a single-dimensional or multidimensional array, depending on what kind of records are being documented. The third is the data package, which allows for the preservation of original files of any format for sharing, compliance, and preservation reasons. The fourth is data provenance and fixity, which tracks any changes that have been made to the data.[29]

Fit-gap analysis

To understand which existing formats best fulfill SRDA requirements, we conducted a fit-gap analysis. We derived a list of requirements from the pharmaceutical use cases described earlier that we used to compare the different formats. The requirements were grouped into three classes: (i) business requirements to represent and implement important data models; (ii) technical requirements to deliver scalable SRDAs; and (iii) application requirements to ensure the usability of SRDAs for processes and

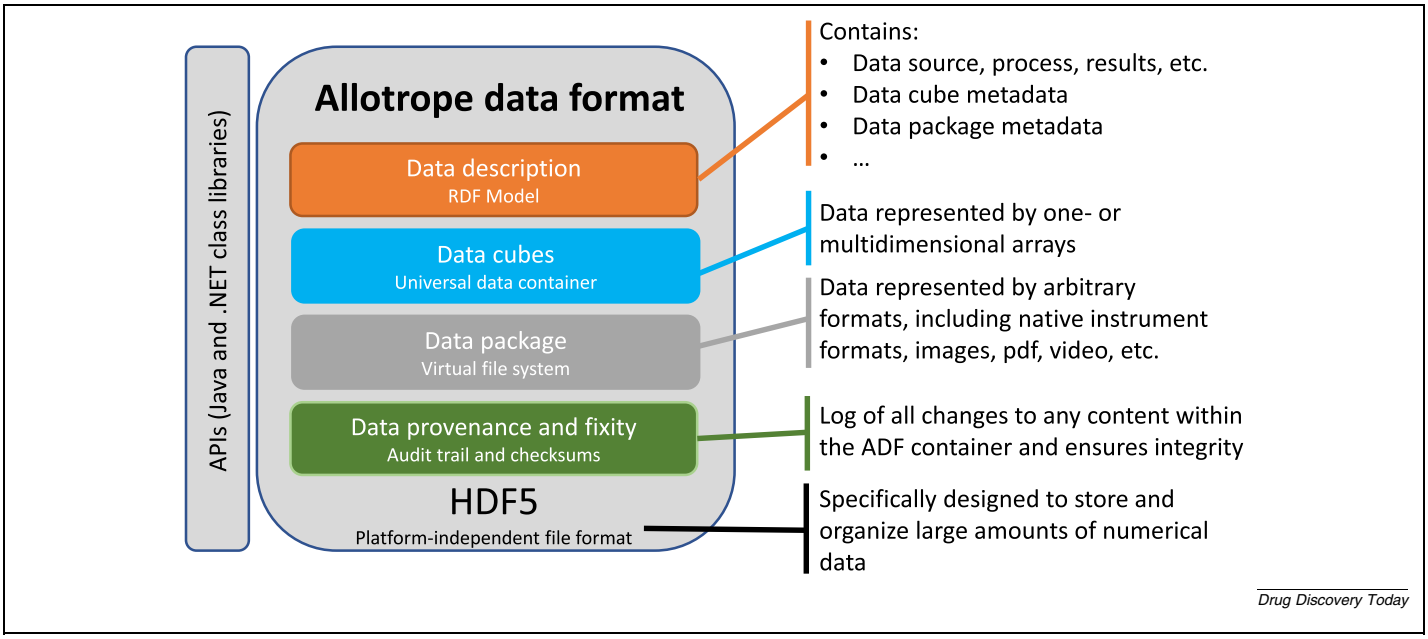

FIGURE 1
Illustration of the structure of an Allotrope data format (ADF) file.

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

analytics. The following sections present the details of this comparison.

Content requirements

Completeness of contextual metadata

To support a consistent understanding of digitally preserved data, the metadata that provide fundamental contextual information must be sufficiently complete for the intended purpose. Unfortunately, JCAMP-DX offers no mechanisms for enforcing sufficient completeness when preserving metadata. For JSON and AnIML, the use of schemas partially ensures that metadata are captured, but this mechanism still allows for the creation of blank metadata structures. By contrast, ADF provides a logical metadata model and enables metadata to be provided in their entirety for reuse and long-term preservation. An SRDA with a fully described context allows for rapid interpretation and reuse by scientists.

LC/UV data models

Similarly, other instruments produce data that follow LC-UV models. JCAMP-DX and JSON do not offer any support to these models, whereas AnIML partially supports some of them. Meanwhile, ADF offers full support to these models.

Mass spectrometry data models

Yet another subset of instruments produces information according to mass spectrometry models. JSON does not offer support to these models, whereas AnIML and ADF partially support them. A working group is actively trying to close this gap for ADF. By contrast, JCAMP-DX offers full support to these models.[30]

NMR data models

A certain subset of instruments produces information according to a NMR data model. Whereas JSON does not offer any support to this model, JCAMP-DX, AnIML, and ADF partially support it (i.e., the latter two offer some technique definitions and ontologies for NMR, but not all required components are publicly available).

Responsive standardization body or community

Although not directly derived from a use case, a reasonable requirement for any standard is an active community that ensures consistent updates and developments to maintain and adjust a format over time. JCAMP-DX lacks an official support group and is only maintained by a small group of developers, most with direct association with vendor company, Bruker. JSON lacks an official standardization body and, although some companies build solutions around it, it does not qualify as a standard. AnIML is supported by a small community with a limited number of companies that drive it forward. The current specification was released in 2007 as a draft and has not been updated since. Whereas JCAMP-DX and JSON do not fulfill this requirement, AnIML does so partially. By contrast, ADF is maintained by the well-established Allotrope Foundation, with over ten large pharma member companies and over 50 partner companies, including major laboratory and software vendors, consulting firms, and academic groups. Monthly community calls, workings groups, full-time employees, and biannual workshops within the Allotrope Foundation completely fulfill this requirement. A scientist with needs currently unmet by existing standards is most likely to find support within the Allotrope Foundation.

Standardized data model support

Scientific instruments often produce data by following a standardized data model. Using a data format that is compatible with the instrument simplifies the reuse, long-term preservation, and exchange of information. As a general format, JSON offers no support to standardized models. JCAMP-DX partially fulfills this requirement, but only for a narrow range of instruments. By contrast, AnIML and ADF fully support the use of standardized data models. ADF especially benefits from multiple active working groups, who have released over 30 data models so far and are actively developing additional models.

Technical requirements

Ease of generation

A cornerstone of data centricity is the standardization of data structures, which is also a goal of SRDAs; one of the ways of fulfilling this requirement is by generating files in a consistent format. This generation process should be easy to encourage consistent data structure. For ADF, manual file generation is not easy because this format has a complex structure and requires multiple layers of information. However, multiple solutions exist that programmatically create high-quality ADF files. For JCAMP-DX and AnIML, the manual process is more straightforward but still requires the data to abide by specific schemas, which add to the complexity of the process. For JSON, this requirement could be completely fulfilled, because this format is flexible and does not require any schemas or additional layers to be observed. However, without a schema, a JSON file lacks standardization and, therefore, only partially meets this requirement. The authors recommend using a schema when considering JSON as an SRDA candidate. Scientists should focus less on manually creating SRDAs but rather on deploying high performance systems to automate data generation, which is easiest for standardized formats, such as ADF.

Efficiently extensible

When centralizing data with SRDAs, it is important to create structures that can be easily extended with additional information, thus preserving a consistent data model. A concrete example would be adding new chromatographs to an existing report inside an SRDA or to expand the list of descriptive metadata tags. This is not possible for JCAMP-DX, because files in this format require decoding before modification. For AnIML, extension is relatively efficient, although it must strictly observe the structure imposed by the schema. One limitation of AnIML is that the metadata reference and data model are inextricably bound up in the XML schema. Given that AnIML cannot leverage an external source, any updates to, or addition of, metadata will require a version change to the format. By contrast, ADF uses a modular architecture to address this point, which enables efficient extensibility. Once again, JSON and ADF fully support this requirement because these formats are flexible and allow for extension.

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

Efficient storing and input/output of binary data

Despite reducing costs of data storage, the efficient storing of binary data is crucial to allow for performant operations on SRDAs. JCAMP-DX, AnIML, and JSON are text-based formats, which leads to large files for binary data, even when encoded as base64, and makes reading/writing slow for large-volume binary data, such as high-end mass spectrometry data. By contrast, ADF is based on HDF5, which stores binary data in a compact format and supports fast input/output of binary data. The scientist benefits from reduced cloud-hosting costs and faster access to relevant measurement data from efficiently stored binary data.

Flexible for many sources

It is also fundamental to unify information from many different sources under a consistent data structure, allowing for the unification of previously disparate assets. Unfortunately, JCAMP-DX does not support this requirement, because this format is only suitable for a specific type of scientific data. AnIML partially meets this requirement, as long as the data fit the predefined format structure. JSON and ADF fully support this requirement, seamlessly handling data from varied domains.

Inbuilt audit trailing

One of the advantages of digital preservation is that it allows users to view the history of data, from their provenance to their current state. Audit trailing is especially important to guarantee GxP compliance and data genealogy. Although JCAMP-DX and JSON do not provide automated audit trail capabilities, AnIML partially provides it by incorporating it as an element of its schema. By contrast, ADF offers fully automated audit trailing, providing a complete record for the use of this functionality.

Inbuilt checksums

To detect errors resulting from the transfer or corruption of data after long-term digital preservation, checksum mechanisms can be used to verify data integrity. Although JCAMP-DX, JSON, and AnIML do not offer integrated and API-supported checksum functionality, ADF does. Instead, JCAMP-DX offers limited error checking of raw data in some cases and AnIML supports signatures over instance data, which partially fulfill this requirement.

Long-term stable format

A key requirement for SRDAs is that the underlying format can digitally preserve data in a stable manner, with a long-term outlook. Although JCAMP-DX has been established since the late 1980s, it lacks detailed documentation and does not have a stable source of financial support. JSON is a flexible format that is rapidly changing, and this, combined with its lack of strict guidelines for archival, means that data might become more difficult to access the longer they are preserved. AnIML relies on its underlying XML standard, which requires users to retain this format across the organization when preserving and transporting data. ADF uses semantic layers, capturing contextual information about the data and making them stable for reuse and long-term preservation. Without using a long-term stable format, scientists risk that their data will no longer be usable in a few years.

Random read/write access

A nonfunctional requirement that encourages data centricity and the implementation of SRDAs is the ability to nonsequentially read or write sections of data, because this allows information to be quickly accessed and, thus, preserved in a unified way. A concrete example would be reading out one pair of X and Y values from a large set of chromatographs stored in the same SRDA, without first parsing the full file. Whereas JCAMP-DX, JSON, and AnIML were not designed to allow random file accesses, ADF fully supports it.

Application requirements

CRO data exchange

It is crucial for data formats to facilitate collaboration between CROs and pharmaceutical companies. Although JCAMP-DX, JSON, and AnIML are formats that offer some data standardization, ADF is the only format that offers separate layers of semantic information that fully explain and contextualize the data.

Ease of building GxP validated workflows

When processing information with the intent of structuring it in a unified and consistent manner (i.e., supporting data centricity and the creation of SRDAs), it is fundamental to abide by industry guidelines and regulations (Good Laboratory, Clinical, Manufacturing, or other Practices; GxP). These requirements are met by designing suitable workflows for data processing, and the choice of a suitable data format is a crucial component in this system. Although JCAMP-DX and JSON can be made to abide by GxP, they were not inherently created to address this requirement. By contrast, AnIML and ADF fully cater to GxP compliance because they have dedicated audit trail functionality to ensure data lineage through the information lifecycle.

Ensures data quality for data sharing or merger and acquisitions

Another reality in the biotech and pharma industries is that companies often collaborate or are consolidated through mergers and acquisitions, and part of this process involves the combination of independent data collections into a single high-quality set. JCAMP-DX and JSON were not designed to support this process. For AnIML, its use of schemas and audit trails provides some robustness when merging information. Finally, ADF can fully support this process, offering audit trailing and layers of metadata that provide contextual information about the data in each subcollection.

Findability through metadata

Being able to find and access data by searching its associated metadata information is crucial for digital preservation and data reuse. JCAMP-DX, JSON, and AnIML partially support this requirement, providing mechanisms for defining data schemas that capture syntactic information. However, these formats do not provide a way of capturing the semantic definitions associated with this metadata. Meanwhile, ADF has a layer dedicated to describing the semantic model of the data, providing a robust way of capturing specific metadata.

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

Interoperability with other techniques

In the context of biotech and pharma, it is crucial to ensure that systems have a high degree of interoperability, which is largely dependent on the formats used for data reuse, long-term preservation, and exchange. This interoperability can be defined either at the physical level or at the data content level, and this discussion applies to the latter. In this context, JSON provides no interoperability (because it has no standardization for labeling in the absence of accepted data model schemas). JCAMP-DX and AnIML partially provide interoperability, as long as the same technique definition is used by all parties. ADF is the only format that provides full interoperability, even for multiple technique definitions.

Power holistic data visualization

In the context of data science, it is valuable to structure information in a way that can be easily visualized. Although it is possible to visualize the structure of data in JCAMP-DX, JSON, and AnIML, ADF also offers the ability to visualize the additional contextual metadata structure. With the right tool, a data-centric lab can be rapidly enabled when data are stored in the ADF.[31]

Support machine learning

Another popular demand in the field of data science is the ability to use machine-learning techniques to identify significant patterns in the data. Once again, although the JCAMP-DX, JSON, and AnIML formats do not preclude further analysis using machine-learning algorithms, ADF provides metadata to assist the exploration of information from a machine-learning perspective.

Concluding remarks

A summary of the fit-gap analysis is presented in Table 2, which lists each SRDA requirement considered, its associated industry use case, and how well the different data formats meet it. As shown in Table 2, ADF fully meets most of the requirements, with the exception of the requirements on compatibility with specific data models (partially met), and the requirement on ease of generation (partially meet). By contrast, JCAMP-DX, JSON, and AnIML mostly only partially meet the requirements. Even though JSON fully meets the more general requirements related to flexibility, extensibility, and interoperability, it does not cater to the specific needs of the pharmaceutical industry (e.g., support of specific data models). From this analysis, it becomes clear that ADF is the data format that best meets the requirements for SRDAs.

In conclusion, this work derived requirements for SRDAs from common pharmaceutical industry use cases and analyzed the ability of four scientific data formats to fulfill them: JCAMP-DX, JSON, AnIML, and ADF. More specifically, it compared these four formats to determine which could best meet the requirements for the concept of SRDAs, which correspond to data that are preserved in a way that also captures contextual information. A fit-gap analysis was carried out to determine how well each format meets SRDA requirements, and results showed that ADF was the format best suited to the needs of the pharmaceutical industry. Beyond the evaluated data models, ADF supports currently over 30 additional data models, with more being released annually by an active community. Only ADF files following the standard in full and using the Allotrope Data Models qualify as SRDAs. Simply using ADF as a wrapper around JSON or other files

TABLE 2

Which existing format fulfills SRDA requirements best?

Requirement	Use case	Format			
		JCAMP-DX	JSON	AnIML	ADF
Content requirements					
Completeness of contextual metadata	**Digital preservation**	**No**	**Partial**	**Partial**	**Full**
LC/UV Data models	Instruments	No	No	Partial	Full
Mass spectrometry data models	Instruments	Full	No	Partial	Partial
NMR data models	Instruments	Partial	No	Partial	Partial
Responsive standardization body or community	Long-term support	No	No	Partial	Full
Standardized data model support	Instruments	Partial	No	Full	Full
Technical requirements					
Ease of generation	Data centricity	Partial	Partial	Partial	Partial
Efficiently extensible	Data centricity	No	Full	Partial	Full
Efficient storing and input/output of binary data	Data science	No	No	No	Full
Flexible for many sources	Data centricity	No	Full	Partial	Full
Inbuilt audit trailing	Digital preservation	No	No	Partial	Full
Inbuilt checksums	Digital preservation	Partial	No	Partial	Full
Long-term stable format	Digital preservation	Partial	No	Partial	Full
Random read/write access	Data centricity	No	No	No	Full
Application requirements					
CRO data exchange	Collaboration	Partial	Partial	Partial	Full
Ease of building GxP-validated workflows	Data centricity	Partial	Partial	Full	Full
Ensures data quality for sharing/M&A	Biotech and pharma	No	No	Partial	Full
Findability through metadata	Digital preservation	Partial	Partial	Partial	Full
Interoperability with other techniques	Biotech and pharma	Partial	No	Partial	Full
Power holistic data visualization	Data science	Partial	Partial	Partial	Full
Support machine learning	Data science	Partial	Partial	Partial	Full

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019

is not sufficient and leads to files that do not pass ADF validation. In the future, ADF should be analyzed according to FAIR data principles to determine whether it is similarly well suited to the needs of a broader variety of industrial domains in addition to pharmaceuticals.

Declaration of interests

D.E.V. is a former board member of Allotrope Foundation. D. D.C., W.C., and A.S.d.S. are affiliated with ZONTAL, which sells commercial products leveraging JSON and Allotrope Data Format. H.F. is affiliated with Agilent, which has a commercial product leveraging the Allotrope Data Format. This manuscript is published independent of Allotrope Foundation and the statements and conclusions of the authors are their own.

Acknowledgments

We thank Ralph Haefeli, Monika Berrondo, James Blair, and Ralph Mueller for inspiring discussions and insights into their respective workflows and challenges. We thank Torsten Osthus for a critical review of the manuscript and helpful suggestions.

References

1 T. Schultz, Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle, Bulletin of the American Society for Information Science and Technology 39 (5) (2013) 34–40.
2 J. Wise, A.G. de Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, et al., Implementation and relevance of FAIR data principles in biopharmaceutical R&D, Drug Discovery Today 24 (4) (2019) 933–938.
3 J. Wise, A. Möller, D. Christie, D. Kalra, E. Brodsky, E. Georgieva, G. Jones, I. Smith, L. Greiffenberg, M. McCarthy, M. Arend, O. Luttringer, S. Kloss, S. Arlington, The positive impacts of real-world data on the challenges facing the evolution of biopharma, Drug Discovery Today 23 (4) (2018) 788–801.
4 V. Steinwandter, D. Borchert, C. Herwig, Data science tools and applications on the way to Pharma 4.0, Drug Discovery Today 24 (9) (2019) 1795–1805.
5 M. Boeckhout, G.A. Zielhuis, A.L. Bredenoord, The FAIR guiding principles for data stewardship: fair enough?, European Journal of Human Genetics 26 (7) (2018) 931–936
6 A. Rodríguez-Iglesias, A. Rodríguez-González, A.G. Irvine, A. Sesma, M. Urban, K. E. Hammond-Kosack, et al., Publishing FAIR data: an exemplar methodology utilizing PHI-base, Frontiers in Plant Science 7 (2016) 641.
7 H. Weigand, P. Elsas, Auditability as a design problem, IEEE 2019 (2019) 276–285.
8 L. Zhou, A. Fu, S. Yu, M. Su, B. Kuang, Data integrity verification of the outsourced big data in the cloud environment: a survey, Journal of Network and Computer Applications 122 (2018) 1–15.
9 R.S. McDonald, P.A. Wilks, JCAMP-DX: a standard form for exchange of infrared spectra in computer readable form, Applied Spectroscopy 42 (1) (1988) 151–162.
10 Pezoa F, Reutter JL, Suarez F, Ugarte M, Vrgoč D. Foundations of JSON schema. In: WWW '16: Proceedings of the 25th International Conference on World Wide Web. Geneva; International World Wide Web Conferences Steering Committee; 2016: 263–273.
11 B.A. Schäfer, D. Poetz, G.W. Kramer, Documenting laboratory workflows using the analytical information markup language, Journal of the Association for Laboratory Automation. 9 (6) (2004) 375–381.
12 A. Roth, R. Jopp, R. Schäfer, G.W. Kramer, Automated generation of AnIML documents by analytical instruments, Journal of the Association for Laboratory Automation. 11 (4) (2006) 247–253.
13 Millecam T, Jarrett AJ, Young N, Vanderwall DE, Della Corte D. Coming of age of allotrope: proceedings from the Fall 2020 Allotrope Connect. Drug Discovery Today. Published online April 5, 2021. https://dx.doi.org/10.1016/j.drudis.2021.03.028
14 Koranne S. Handbook of Open Source Tools. Boston: Springer; 2011.
15 Oberkampf H, Krieg H, Senger C, Weber T, Colsman W. Allotrope Data Format – Semantic Data Management in Life Sciences. Published online December 5, 2018. http://dx.doi.org/10.6084/m9.figshare.7346489.v1
16 A. Hasnain, D. Rebholz-Schuhmann, Biomedical semantic resources for drug discovery platforms, Lecture Notes in Computer Science 10577 (2017) 199–218.
17 J. Peckham, F. Maryanski, Semantic data models, ACM Computing Surveys 20 (3) (1988) 153–189.
18 L.A. Treinish, M.L. Gough, A software package for the data-independent management of multidimensional data, Transactions American Geophysical Union 68 (28) (1987) 633–635.
19 R. Rew, G. Davis, NetCDF: an interface for scientific data access, IEEE Computer Graphics and Applications 10 (4) (1990) 76–82.
20 Lee CA. Open archival information system (OAIS) reference model. In: Bates MJ, Maack, MN, eds. Encyclopedia of Library and Information Sciences (3rd edn). London; Taylor & Francis: 2010; 4020–4030.
21 Katz P, Campbell C. FDA 2011 Process Validation Guidance: process validation revisited. Journal of GXP Compliance 2012; 16(4): XXX–YYY.
22 P. Campbell, Data's shameful neglect, Nature 461 (7261) (2009) 145.
23 C. Frisz, G. Brown, S. Waggoner, Assessing migration risk for scientific data formats, International Journal of Digital Curation 7 (1) (2012) 27–38.
24 Bourhis P, Reutter JL, Suárez F, Vrgoč D. JSON: data model, query languages and schema specification. arXiv arXiv:1701.02221v1.
25 Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F. Extensible markup language (XML) 1.0. W3C recommendation. https://www.w3.org/TR/xml/ Accessed July 18, 2021.
26 M. Mani, D. Lee, R.R. Muntz, Semantic data modeling using XML schemas, Lecture Notes in Computer Science 2224 (2001) 149–163.
27 AnIML. An AnIML document. https://www.animl.org/an-animl-document Accessed July 18, 2021.
28 S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks, The semantic web: the roles of XML and RDF, IEEE Internet Computing 4 (5) (2000) 63–73.
29 D. Della Corte, W. Colsman, B. Welker, B. Rennick, Library eArchiving with ZONTAL space and the allotrope data format, Digital Library Perspectives 36 (1) (2020) 69–77.
30 P. Lampen, H. Hillig, A.N. Davies, M. Linscheid, JCAMP-DX for mass spectrometry, Applied Spectroscopy 48 (12) (1994) 1545–1552.
31 D. Della Corte, K.A. Della Corte, The Data-Centric Lab: A Pharmaceutical Perspective, Advances in Intelligent Systems and Computing 1364 (2021) 1–15, https://doi.org/10.1007/978-3-030-73103-8_1.

Please cite this article in press as: D. Della Corte et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.07.019