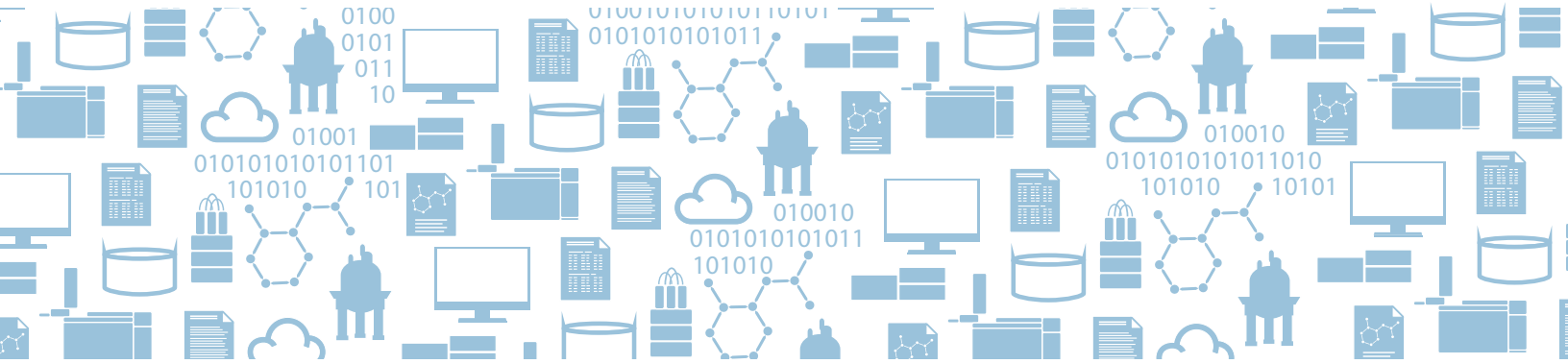
White Paper

Looking Beyond Analytical Data Standardization—the Fourth Paradigm



Co-authors: Andrew A. Anderson, Graham A. McGibbon, Andrey Paramonov, and Sanjivanjit K. Bhal

ACD/Labs

Executive Summary

Chemical R&D activities continue to generate a deluge of instrumental analytical data on a daily basis, regardless of industry. Regulatory submissions and **critical R&D or manufacturing decisions are based on analytical data every day**. When data are silo'd and unavailable in standard accessible formats, access and re-use for decision-making and problem-solving is hard, if not impossible. **Organizations must have ways to standardize, homogenize, and digitize analytical data to improve data access while maintaining data integrity**, and facilitating scientific business innovation. In this drive for standardization, however, we postulate the importance of positioning it with chemical context, since analytical experiments are diverse by purpose and support many chemical workflows.

Homogeneous, assembled, digitized analytical data lends itself to be included in the stream of meaningful data exchange between external organizations and data-sharing inside organizations. This organization and data transformation is necessary to effectively build the 'data-to-information-to-knowledge' lineage that enables managers to make strategic and tactical decisions; to maximize benefits, and limit risks. This paper provides commentary on efforts for analytical data standardization and a vision of the broader considerations and requirements that such an undertaking should include.

Introduction: Data-to-Information-to-Knowledge

In ca. 1600 AD, Johannes Kepler published new interpretations of empirical scientific data to postulate new insights into how the universe worked. In the foreword to the 2009 seminal text, 'The Fourth Paradigm—Data-Intensive Scientific Discovery' (inspired by Jim Gray's 2007 paper on eScience), Gordon Bell from Microsoft Research relays the following:

"It was Tycho Brahe's assistant Johannes Kepler who took Brahe's systematic astronomical observations and discovered the laws of planetary motion. This established the division between the mining and analysis of captured and carefully archived experimental data and the creation of theories."[1]

In his influential paper 'eScience: A Transformed Scientific Method' Jim Gray described Data Exploration as the fourth paradigm:

"A Thousand years ago, science was empirical, describing natural phenomena—[first paradigm]

In the last few hundred years, the theoretical branch used models and generalizations—[second paradigm]

In the last few decades, the computational branch afforded the simulation of complex phenomena—[third paradigm]

Today, data exploration (or eScience) aims to unify theory, experiment, and simulation—[fourth paradigm]:

- *Data is captured by instruments or generated by a simulator*
- *Data is processed by software*
- *Information/knowledge is stored in a computer*
- *Scientist analyzes database/files using data management and statistics"* [2]

In this fourth paradigm data is, therefore, the lineage of information which provides knowledge that enables managers to make strategic and tactical data-based decisions

for actions that maximize benefits and limit risks. Data exchange between organizations and data sharing inside organizations is necessary to effectively communicate this 'data-information-knowledge' lineage. Such an approach, however, demands dealing with the data deluge coming from the overwhelming volume, velocity, variety, and variability of data. This is no truer than when discussing analytical data generated on a variety of instruments; from disparate techniques; to answer any number of different questions and address diverse situations throughout chemical R&D.

Data Sources: Heterogeneity and Analytical Chemistry Data

In data workflows, not only big data but analyses generating a variety of not-quite-so-big data, two factors contribute greatly to the deluge. The first is the automation and/or parallelization of specific high-throughput analyses on particular instruments. The second is the challenging implementation of the so-called 'Internet-of-things (IoT)' due to the tremendous assortment of computer-based data sources and their diversity of parts, performance attributes, and output of *analytical* data formats. All that further highlights the benefits that are to be gained by automatic, on-the-fly digitizing of the instrumental analytical data and integrating this data type with chemical data representations. Not only does it create a reliable, free of manual error, comprehensive body of knowledge for decision-making, sharing within and outside the organization, or regulatory submissions; but also 'future-proofs' the laboratory[3] and contributes this knowledge, upon demand, to the organization's overall (big and small) analytics.

Nestled between pen and paper of the dwindling past and the hoped-for Star Trek inspired tricorders of the future, is a cornucopia of proprietary data constructs. This is necessary for data source or instrument hardware innovation and the efficient acquisition and storage of bytes-to-petabytes of analytical data. Ongoing heterogeneity of *instrumental analytical chemistry* formats is thus a natural hallmark of technology advancement. Naturally though, having many different data formats creates a desire for standardization. There are, however, different motivations behind constructing a so-called ideal standard (free or not; encoded vs. readable; open source or closed source).

There are two approaches for dealing with heterogeneity. The first, favored by many data scientists, is to force a single, all-encompassing format. The second is to be able to exchange data and information between formats. Being able to do both is also a possibility. A consideration of the merits of open vs. closed and free vs. proprietary is beyond the scope of this paper, but certainly warrants additional dialog. Effective standardization is not simply the same as having human parsable data—as purely open generic formats such as ASCII text or XML exemplify. It is reasonably simple to establish standards for single, homogeneous datatypes. For example, human-audible audio data formats were consolidated and standardized in the early 2000s, however, video streams are standardized in a different way. Given the wide variety of analytical data types, sizes, and content, creation of a single standard has been a continuous challenge.

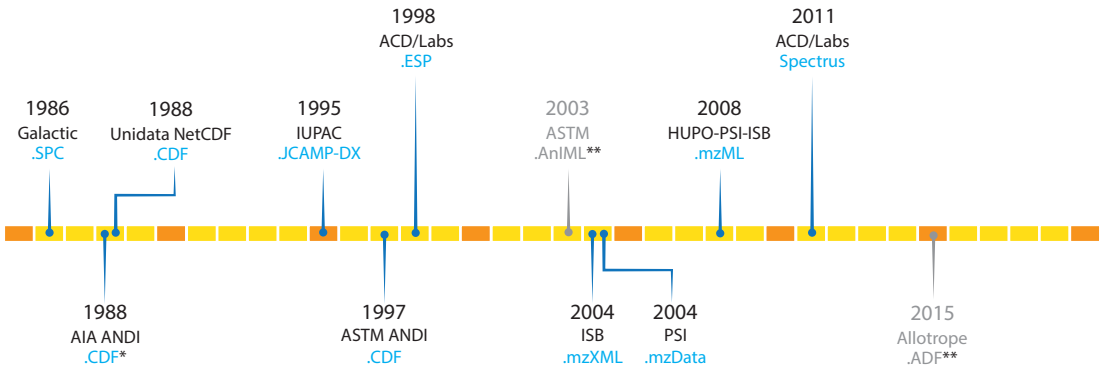
It should be noted that while humans still need to access data in some form, it is increasingly important to have formats that are efficient for digital manipulations. Long ranging benefits of digitization include:

- Minimizing manual human intervention within a multistep process from experiment planning, through data acquisition and analysis, to result delivery

- Enablement of the **assembly of analytical data** into projects that support chemistry decision-making such as analysis, discovery, confirmation, verification, identification, and controls
- Support for future projects through efficient data accessibility
- Inclusion of analytical chemistry to the stream of growing data in the modern chemical enterprise for access, searching, and decision-making; but also for the enterprise-wide business analytics

Paradigm Pre-requisite: Analytical Data Standardization and Digitization

For a chemistry driven organization's day to day operation, **digitizing analytical chemistry** enables broader access, more comprehensive analysis, and assembly. This benefits both each scientist in the lab (remote or on premises) and decision-makers elsewhere in the organization. Analytical data are generally used for qualitative (*what is in my sample?*) and quantitative (*how much of each analyte is in my sample?*) investigations. Depending upon sample composition and the physical characteristics or properties of the analytes of interest, a wide range of techniques may be used to gather analytical data. Among the most common are separations, chromatographies for small molecules (LC, GC, SFC) and electrophoresis for biomolecules (PAGE), mass spectrometry, and spectroscopies (NMR, UV, IR, CD, Raman). Typically, each vendor creates its own proprietary format for data acquisition and data handling.



*1992–updated for chromatographic data, 1994–updated for MS data **In Beta

Figure 1 Notable analytical data standardization efforts†

†This image is not intended as an exhaustive summary. Some of the efforts are covered further in the document

One of the earliest efforts for analytical file format standardization was by the Galactic Company starting back in 1986, with a binary format for a variety of spectroscopic data (SPC). Also in the 1980s, the ASTM E01.25 sub-committee for Laboratory Analytical Data Interchange protocols and Information Management were conducting an effort that led to the ANDI data standard (NetCDF).[4] Circa 1995, the Joint Committee on Atomic and Molecular Physical Data format (JCAMP-DX)[5] was established with the involvement of the International Union of Pure and Applied Chemistry (IUPAC). In 2003 IUPAC was looking toward markup languages for data formats. Not surprisingly, the same year the ASTM E13.15 subcommittee with the involvement of a range of stakeholders spearheaded

an initiative to make a new XML-based standard (AnIML)[6] for analytical data. During its emergence, the mass spectrometry community, among others, were also actively trying to create other exchangeable analytical data formats (mzXML, mzData), and eventually the de facto standard mzML evolved.[7] More recently a subset of organizations from the pharmaceutical industry formed the Allotrope Foundation with an aim to establish an Analytical Data Taxonomy (ADT) and also to create their own Allotrope Data Format (ADF),[8] contracting a third-party software company to build a supporting software framework. Like others before, prior experiences are being considered but making the ADT seems the newer, interesting part of the exercise.

On the **proprietary format standardization** front, ACD/Labs has been actively amassing import capabilities of analytical instrument vendor formats within their own proprietary format, since 1998. The current *.spectrus format covers the majority of spectral and chromatographic formats in laboratory use; as well as the most popular open formats (ASCII, JCAMP); and evolving standards such as the Allotrope Data Framework (ADF). Major instrument vendors (e.g., Agilent, Bruker, LECO, PerkinElmer, Sciex, Shimadzu, Thermo, Waters, etc.) keep their data formats closed but provide software development kits (SDKs) for data access by third parties, and enable export in some of the above-mentioned standard formats.

Some of the initiatives to supersede older data standards—such as AIA/netCDF format—are undertaken to gain value from the assembly of multiple types of data. Some can be attributed to dreams of big data science, while others are due to the complexities and innovations in analytical equipment, especially mass spectrometry. Mass Spectrometry (MS) commonly generates the most complex and largest size datasets compared to other detectors. As an additional complication, experiments often simultaneously include those additional data dimensions or types, e.g., ion mobility, or imaging MS. MS hardware technologies are rapidly developing and there are a wide variety of experiments and workflows. Some examples of MS specific standard formats include: SPC/Institute for Systems Biology .mzXML[7]—an early XML format for '-omics' research; HUPO-PSI .mzML[8]—backed by ProteoWizard implementation (probably the most advanced free MS software at the moment; it was conceived as a single consolidated XML format for MS data instead of several others including .mzXML noted above); and .mz5[9]—a more efficient implementation of .mzML ideas based on .HDF5 (but not actually widely used).

Since the identification of compounds by mass spectrometry is one of the most common qualitative analyses, chemical structure information is increasingly realized as important to assemble with MS data, but definitive structure elucidations inevitably also rely on additional data (NMR, IR, Raman, etc.). For other workflows, different combinations of analytical chemistry measurements are necessary to support the end goal of the investigative process of **molecular characterization**. It is important to note that the true benefit comes from the ability to bring different analytical data in a homogeneous environment, oftentimes paired with the chemical content of the given project, for effective delivery of the results. The truth is that for decades that homogeneous environment has been a physical desk, or lately a Microsoft Office product (Microsoft Word or Excel). This is not only a manual process, but also one that restricts the depth and volume of information available in the **instrumental analytical measurement**. Digitizing efforts are underway, however, and commercial products are available to offer electronic multi-technique, multi-vendor data assemblies without data reduction or abstraction.

The Fourth Paradigm–The Necessity of Digital Assemblies

Digital Chemical Structure Representation

When discussing the advantages of digitizing analytical data, it is important to also discuss the **digital representation of chemical structures**. The majority of analytical experiments are run to help scientists identify and/or characterize discrete chemical entities, mixtures, and formulations. The ability to connect spectral/chromatographic peaks to the structure is the very essence of chemistry decision-making. Enabling this connectivity gets us a step closer to knowledge retention, instead of simple data storage for regulatory purposes.

A starting point in the digital representation of chemical structure can be found on Wikipedia which states:

"There are two principal techniques for representing chemical structures in digital databases

a. *As connection tables / adjacency matrices / lists with additional information on bond (edges) and atom attributes (nodes), such as: MDL Molfile, PDB, CML*
b. *As a linear string notation based on depth first or breadth first traversal, such as: SMILES/SMARTS, SLN, WLN, InChI*

These approaches have been refined to allow representation of stereochemical differences and charges as well as special kinds of bonding such as those seen in organo-metallic compounds."[10]

The Molfile which is now property of Biovia of Dessault Systemes[11] has different versions but is commonly considered a de-facto standard for 'small molecules'. An SDfile is just a sequence of Molfile records and is widely used for the exchange of chemical information. However, an SDfile cannot be used as an underlying format for advanced applications, even for database search, because it lacks random access.

One of the major challenges of representations for biomolecules is the canonical relationship between certain types, namely DNA, RNA, and the amino acid sequences of proteins and peptides. The most popular 'large molecule' (peptides, proteins, DNA, RNA, etc.) formats are: FASTA,[12] PDB,[13] and Hierarchical Editing Language for Macromolecules (HELM)[14]—a format of large molecule representation which falls into the Chemical Structures/Large Molecules niche and is definitely not an all-embracing data format.

Companies and vendors that use chemical structure extensively will generally also use their own storage formats. Within the context of instrumental analytical chemistry, for example, the ACD/Labs Spectrus Platform incorporates and interlinks chemistry representation within **analytical data interpretation**. This **assembled data** may be stored within Oracle or PostgreSQL databases or a proprietary Spectrus *.cfd file.[15]

Different groups with a variety of stakeholders have driven different standards for digitized chemistry representation. Most of those common formats are fairly mature. The InChI format is the latest and probably the one evolving most recently.

Assembled Analytical and Chemical Data

'**Assemblies**' of analytical data from multiple sources, chemical structures, and interpretations as described already, are the so-called Fourth Paradigm. For a large range of scientific experiments, there are a variety of data types which contribute to the understanding of material, disposition, and behavior.

The **interrelatedness** of certain analytical information is especially important. For many

materials/substances, compositional profiling by hyphenated chromatography-mass spectrometry analyses may be captured by one data file. However, more comprehensive characterizations commonly require:

- Data from experiments carried out using multiple techniques from different instruments that generate data in different formats.
- Data that may have been collected and interpreted at different times, in different labs, by people within or external to the scientists' own organization.

For more than two decades ACD/Labs has created software platforms and components that support **digital molecular characterization**. Standardization, using and exchanging proprietary formats (often including automation), has become the cornerstone for meaningfully 'assembled' chemistry data. Members of the Allotrope Foundation, the most recent entrant to the analytical data standardization space, perceive benefits for such software (including their eponymous Framework) as including, "…[to] facilitate regulatory compliance by making data easier to find, visualize and extract knowledge from, throughout its entire lifecycle".[7]

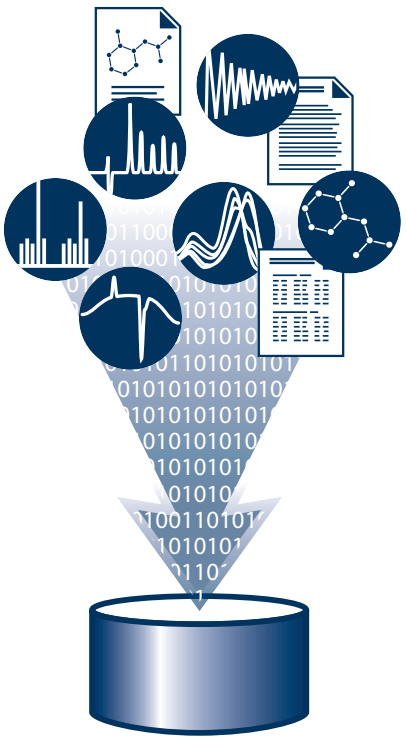
To support interrelatedness, systems in the Fourth Paradigm must be able to not only sufficiently index individual analytical data files, but must also afford 'analysis assembly' capabilities to provide users with a comprehensive 'story' for relevant analyses.

Consider formulation profiling as an example. The following list of 'related data' must be 'assembled' to present a comprehensive assessment of a product formulation:

- LC-UV/MS (and other detector types)
- GC-FID/MS (and other detector types)
- chemical, biological, formulation schematics
- chemical structures for process-related impurities/degradants
- 1D and 2D NMR data for isolated formulation components, with references to separations information (e.g., retention time)
- XRPD, DSC, TGA, particle size distribution, and a variety of other material characterization datasets

Finally, the ability to have explicit digital representations of a scientist's interpretations—specifically beyond alphanumeric descriptions is also necessary in the Fourth Paradigm. In the example of product formulation above, **digital representation of interpretation** would be:

- Chemical structures 'assigned' to spectral and chromatographic components—directly within data architectures.
- Association of experimental metadata to assembled analytical data architectures.



Fulfillment of the Fourth Paradigm

Software platforms that allow the fulfillment of the Fourth Paradigm enable readily accessible digital storage and manipulation of analytical data, chemical structure information, and other meta data. The latest embodiment is ACD/Labs' Spectrus Platform.[15] The 2014 data management platform includes:

- Components capable of reading analytical data from over 130 major vendor proprietary formats (via SDKs and legacy

partnerships developed over two decades) and open formats, for chromatography, spectroscopy and a variety of other XY type data.

- The capability to aggregate instrumental analytical data with chemical and laboratory context for decision-making and further reuse.

Further advancement of standard formats for analytical chemistry will enable greater unity between theoretical science, experiment, and simulation as per the Fourth Paradigm concept. This progression will offer far-reaching benefits to chemistry-driven organizations for **scientific business innovation** and **more efficient commercialization**. We have moved beyond 1600 AD, past the technological advances of the turn of the century, to the age of IBM Watson and cognitive learning/augmented intelligence. Solving the problem of unifying and effectively assembling data is the core prerequisite for advancements that technology will make possible.

References

1. The Fourth Paradigm: Data-intensive Scientific Discovery; Steward Tansley and Kristin Michele Tolle; Microsoft Research, 2009.
2. eScience—A Transformed Scientific Method, Jim Gray, 2009 (https://www.slideshare.net/dullhunk/escience-a-transformed-scientific-method)
3. http://www.allotrope.org/benefits-of-the-framework
4. http://andi.sourceforge.net/
5. http://www.jcamp-dx.org/protocols.html
6. https://www.animl.org/
7. http://www.allotrope.org/
8. http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML
9. http://www.psidev.info/mzml; http://proteowizard.sourceforge.net/
10. https://en.wikipedia.org/wiki/Chemical_database (cf 2016.12.01)
11. http://accelrys.com/products/collaborative-science/biovia-draw/ctfile-no-fee.html
12. http://www.psidev.info/node/363
13. http://www.wwpdb.org/documentation/file-format
14. http://www.pistoiaalliance.org/
15. http://www.acdlabs.com/spectrus/

Andrew A. Anderson
Vice President Innovation, Informatics Strategy, ACD/Labs

Graham A. McGibbon
Director, Strategic Partnerships, ACD/Labs

Andrey Paramonov
Software Developer, ACD/Labs

Sanjivanjit K. Bhal
Director of Marketing and Communications, ACD/Labs