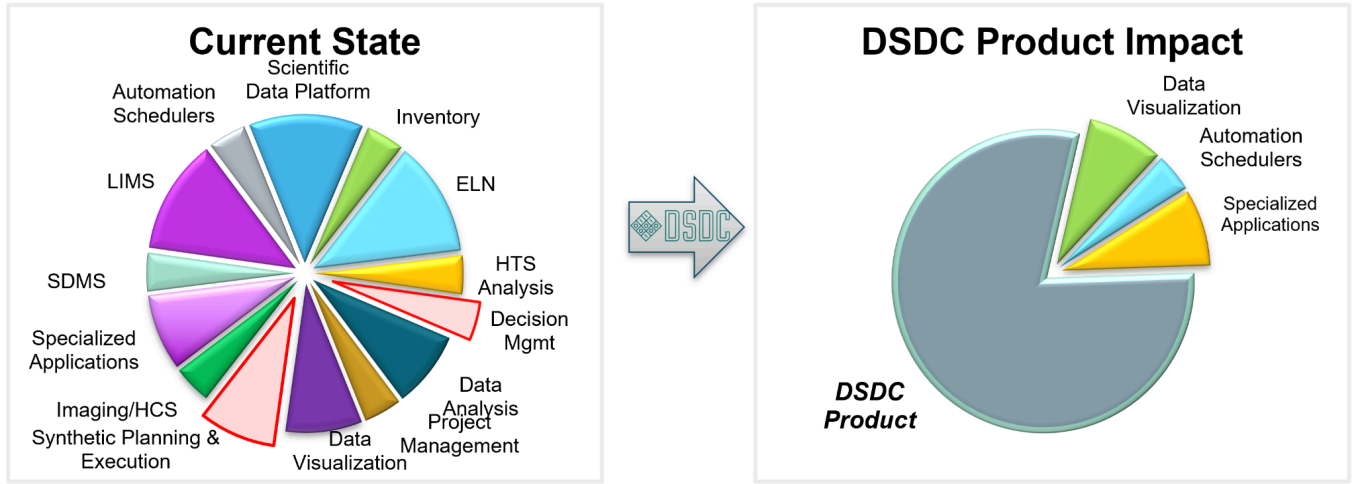


Laboratory Software: Moving from “Inadequately Oversolved” to Enablement

Lately I have been describing the current ecosystem of laboratory informatics as over-solved yet somehow still inadequate. The DSDC seeks to use an industry consortium to solve this problem for ourselves. I will elaborate in another post soon on the details of the consortium itself and the commercial viability and financial modeling which validate this unique approach, but for now I would like to share more about HOW the DSDC product will correct the current state of R&D lab informatics.

Current State - Software to Run R&D Labs

The pie chart on the left is a typical example of the various different commercial and/or home-grown software systems look like in any single given company today. In fact, one will often find this heterogeneous array exists in separate environments (with different administrators and user groups) for each therapeutic area within a single large biopharma company. Most troubling for a user is that most lab scientists and researchers are required to use most if not all of these systems each and every day simply to perform their work - contributing to the 40-60% of non-value added activity recently cited by Accenture at Lab of the Future Congress. Most troubling for the organization who licenses all of these - the cost, averaging around $16MM per year (just for the licenses, nevermind support and services). Some companies are in the $30MM range!

Oversolved

Despite the fact that the fundamental things we use and actions we take in labs are consistent and relatively easy to describe, we have specialized inventory software, ELNs, and LIMS which have a great deal of redundancy between them. Due to the use of many of these systems, over the past 10 years or

so we have seen the advent of “scientific data platforms” & SDMS systems which are necessary to move data from these disparate systems into a single source of FAIR data for the data consumer.

Inadequate

While there are many systems with significant overlapping capability in each ecosystem, I find that the most glaring gaps needed to operate labs efficiently are the tools for synthetic planning and execution. I am not referring to *design* tools - I believe we have a few examples or really great ones - I am referring to the gap between having molecules designed and planning the laboratory activities to actually assemble the resources (e.g. starting materials) and make the molecules. Excel remains the winner here. Additionally, there are no commercial systems aimed at decision management. You may hear the term decision support applied to data visualization applications; however, for a project to create and evolve its knowledge the need to manage the hypotheses and conclusions made in and by that project team exists. Where do we track a decision about why scaffold 4 or vector X was used in lieu of scaffold 2 or vector Y? In Powerpoint.

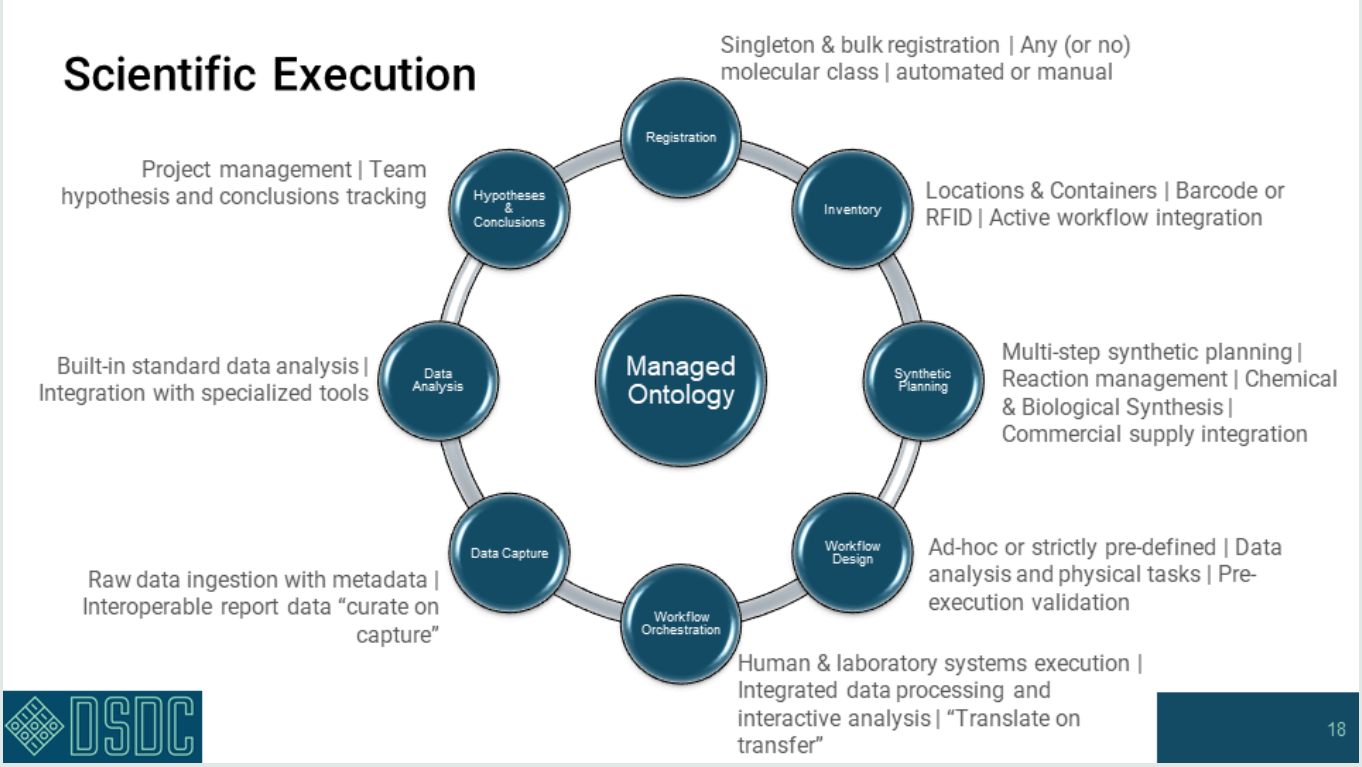
Future State - Enabled Labs

At first glance, most of us who have been “in the trenches” - which means we have baggage - look at the broad array of software and think “there are so many features, no single system can replace them all”. I encourage you to think about what happens in the labs, not what features exist today. At the core, what researchers and technicians do in labs is a very feasible list of activities which require another very feasible list of resources which need to be tracked. As an example, examine this “mostly complete” diagram of informatics product capabilities needed to execute experiments and processes within our labs. Many of the systems used today have many of these capabilities included, in fact, there is a great deal of redundancy. Moving from many systems to one system is actually very achievable so long as that one system is built upon the right technical architecture and has the user interfaces which make work easier, not harder.

Before we get more technical…this is all about prioritizing the user experience

I hesitate to go further into details about the technical details which will enable our future state because that is not the DSDC priority. A technical architecture and novel approaches described below are derived from one objective: An exquisite user experience for our laboratory researchers and technicians. We must have a laboratory informatics infrastructure which makes our job more pleasant, more efficient, and more robust so that we enjoy every day in the lab where we are contributing to medicines and diagnostics which solve health challenges for our communities. The technical details to follow are only valuable in that they are required to facilitate a dramatic improvement to the working lives of research project teams and laboratory operations teams. I would LOVE to put up several amazing user interface designs and user flows which exemplify what a day in the life of a lab user might look like, but this will

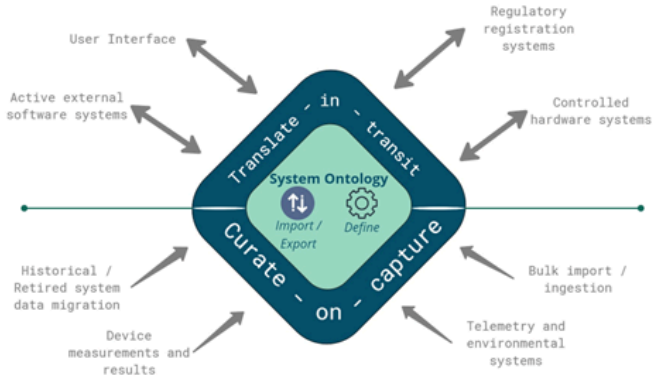
take funding. When the consortium secures investments, rest assured that this product will start with exquisite user interfaces. These user experiences and interfaces; however, don't exist in today's software ecosystem because of constrained underlying technical designs and principles. As I go further with technical details, please remember that these are in service of building a global, indispensable, exquisite user experience for the DSDC product.



At the center of scientific execution lies a mechanism to manage the ontologies used in the many different scientific environments which describe the tasks and materials for scientific execution. Ontologies are important to any industry where many different organizations or stakeholders need to communicate with one another. Often, we get stuck thinking about ontologies as simply a set of standards to name and relate things. While standards can streamline building tools for communicating, they can also have a paralyzing effect when those ontologies (standards) are not agreed upon. In life sciences we have seen many instances of this. Generally, ontologies for new science (~ past 15 years) have been more successful, such as GO (Gene Ontology https://geneontology.org/) or SO (Sequence Ontology http://www.sequenceontology.org/).

The large and heterogeneous drug discovery and development industry-at-large operate businesses which span many different ontologies and many areas without any ontologies. ***The pace of ontologies does not move at the pace of science.*** Consequently, the DSDC product vision includes a new concept of **Dynamic Ontology Management** which is enhanced by ontologies but facilitates their creation or ignoring them altogether.

Ontology Management



Dynamic Ontology Management

- ✓ Configure your system by defining your own ontology through administrative interfaces
- ✓ Import or export ontologies (.owl files) to start, expand, or update your own ontology
- ✓ Interoperability with other systems is managed by configured ontology translators which translate in real-time
- ✓ Translators are exportable and shareable, so common translators (e.g. instrument output) can be easily adapted to your specific ontology

DSDC

The need for widely adopted industry ontologies is negated

21

By shifting the focus from defining a global set of ontologies and building them into a software product, the ontology management for our future state focuses on streamlining the creation and sharing of ontologies within the informatics system which powers our laboratory execution. Building a catalog of translators which actively facilitate data capture and data sharing around the central ontology of any company enables a small number of translators to support a vast ecosystem of bespoke and more broadly recognized ontologies.

From a product strategy lens, this is implemented with users by creating user interfaces to streamlining the import, export, and creation of ontologies. Following that come user interfaces for building and sharing translators and implementing them to “Curate-on-Capture” of your data producers and “Translate-in-transit” for your data consumers which exist around this new software platform. Finally, this is likely to be one of the easiest and most impactful capability areas to benefit from LLMs to dynamically build translators.

This sounds great, but why do we not just do this with one of the many systems which exist today. Well, aside from maybe it just has not been invented prior to this moment, we need to wrangle with the challenge of a data platform which is capable of representing anything and any process for any environment in which we operate labs.

Introducing the Digital Mirror

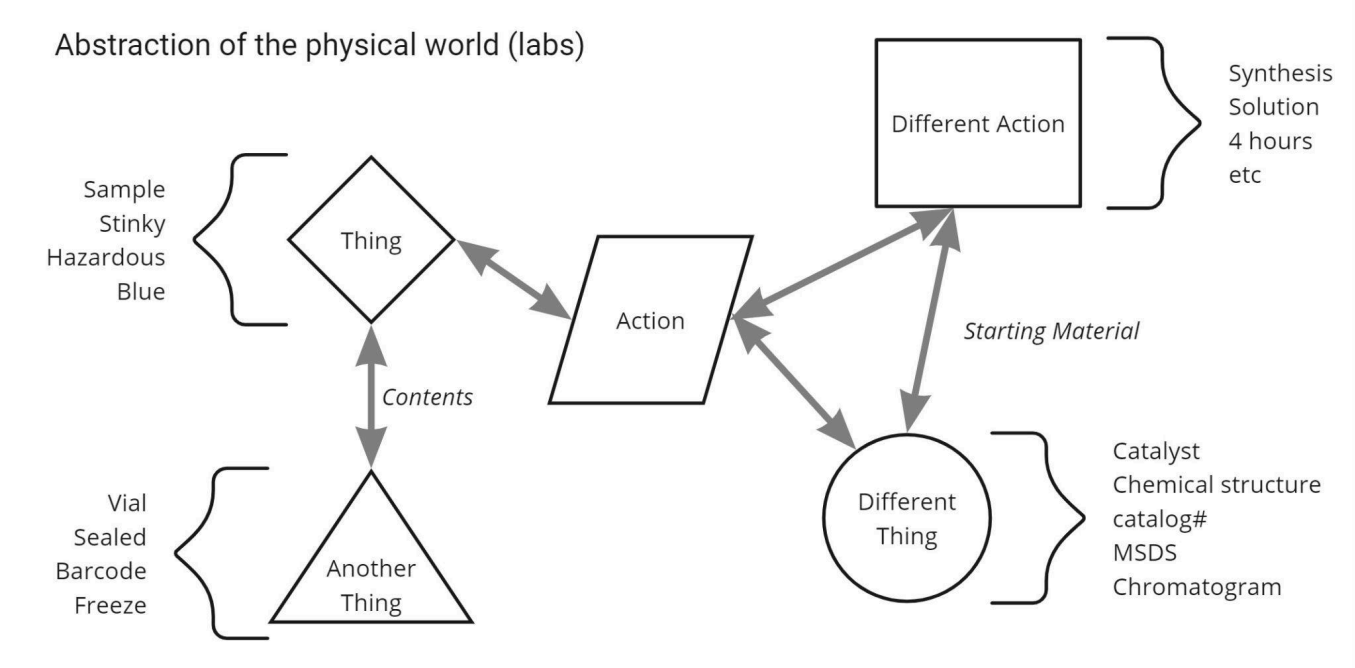
The following has been described to me as highly patentable and protectable. I have been encouraged to “make lots of money” from this. I am publishing this novel architectural pattern so that nobody, including myself, can patent this approach henceforth so that anyone - whether it is myself with the DSDC - or anyone else in the world - can use this to create a true digital mirror of any physical world environment.

There won't be enough details here in my diagram to build it, but the critical elements of a new pattern I call "**Fragmented Persistence Object Service**" (FPOS) model are now public.

PREAMBLE - UNIVERSE ABSTRACTED

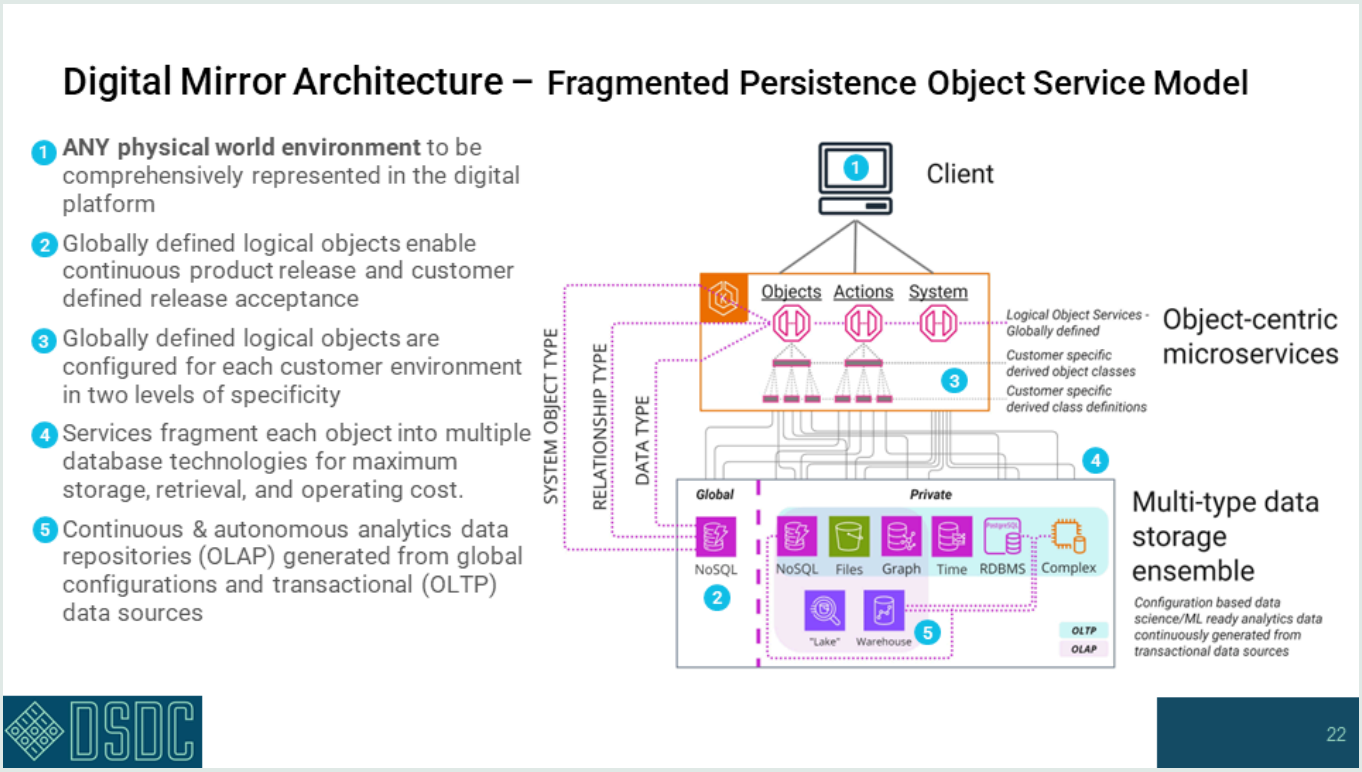
So how do we create a data architecture that never needs to change? It will need to represent any physical environment by only modifying the content (not the structure). The physical world (which btw includes ideas and theories about the physical world we use in science) is very complex and using any given database or data storage technology will inevitably force some perturbations in how we represent the physical world…or at a minimum it will perform poorly in transacting this complex physical environment.

This is solved through abstraction (like so many things in software engineering and mathematics). If we describe an abstract view of the physical laboratory environment, it looks like this:



Simple, right? Think of any part of your life in science. If you keep these super abstract, you won't find a scenario you can't depict. So, let's take this super abstract conceptual model and apply a mechanism to add specificity for any specific environment using a well established (AWS) technology stack.

Climbing deep into that rabbit hole has led me to the FPOS model unveiled here.

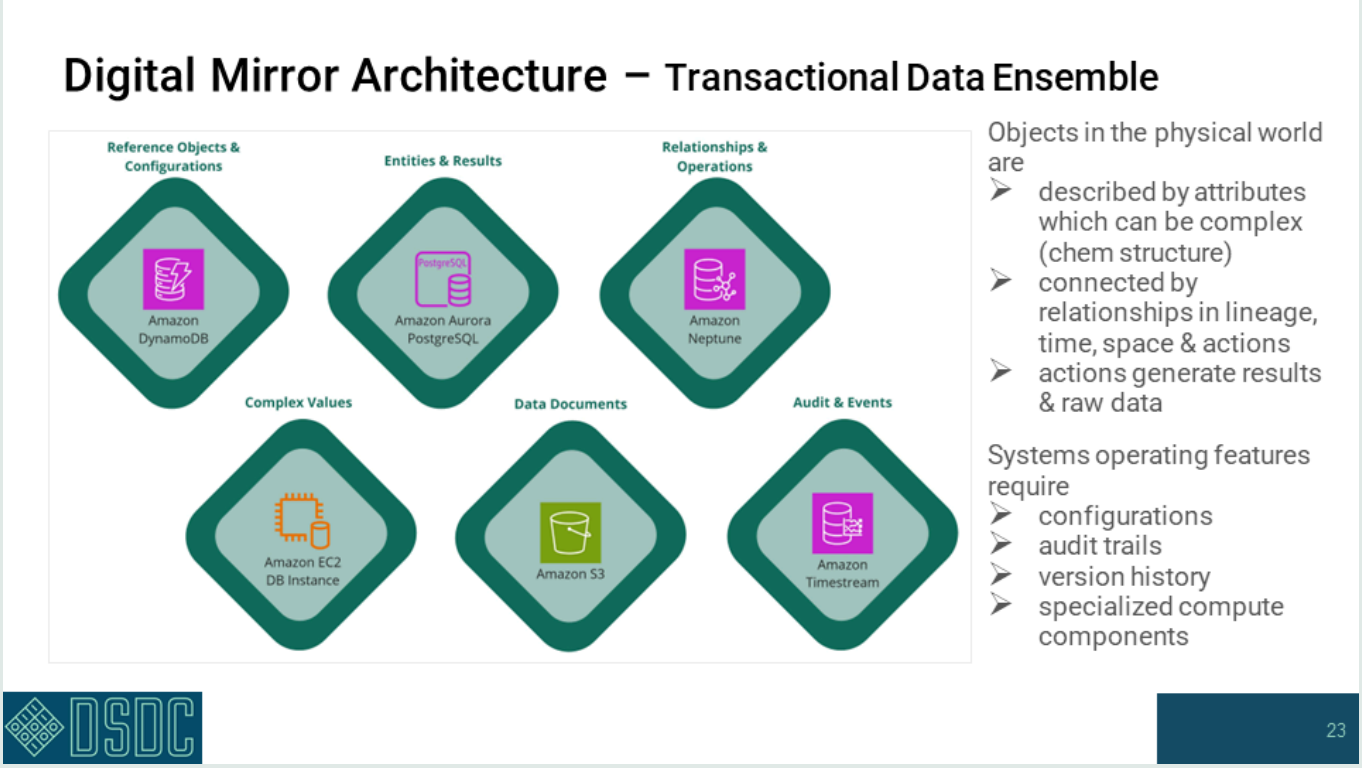


As you step through the slide, you'll note a few key concepts enumerated on the left. There are many technical details buried under the surface of this diagram, most of which a cloud savvy software engineer will easily identify and envision more details of this architecture. For the non-software engineer, here are the critical concepts delivered with this novel architectural pattern:

1. Although it violates the "microservice architecture" pattern, the real world requires that each thing or relationship be represented by an ensemble of different database technologies. The fragmentation of the digital version of a physical thing is critical to be scalable, flexible, and sustainable.
2. The globally defined abstract objects, relationships and data types are tightly coupled to specific versions of the software application, enabling a core set of abstract logical patterns to create highly specific customer environments which can even exist in different versions concurrently.

These days we in the informatics space spend a lot of time and effort on preparing data for consumption. In this digital mirror environment with a centrally managed ontology, the huge obstacles related to preparing data for consumption by data scientists or machine learning platforms becomes a simple task. So simple, in fact, that the ontology configurations you make will also define how native cloud capabilities assemble your data with minimal latency for any variety of consumers. This is all possible because of the unique transactional data ensemble within the FPOS architecture.

The transactional data ensemble where the fragments of data which mirror physical world objects and relationships are the key to ultimately having unrestricted ability to build the single platform which delivers our future state of amazing user experiences. By understanding and utilizing each different database technology to its strengths and NOT using any given DB type for things it is NOT good at, we can use an array of databases to create our digital mirror.



This ensemble encompasses the needs of the platform to function as a software system across a global array of users and organizations AND the ability to fragment any physical world thing or concept into a logic-driven array of fragments. Moreover, these fragments are all stored in the most effective manner possible using the latest native AWS technology.

Conclusion

If you are technical, your head is swimming a bit in the new concepts here. If you are a researcher or technical lab staff, I hope you are still dreaming about what your life might be like if you could have one system for 80% of your needs each day and the entire time it is a pleasant, productive, enriching experience. All of you, please share this with your colleagues. It IS possible to change our global industry laboratory experience.

If you are an executive sponsor or stakeholder, I'll follow up soon with a post which includes the commercial approach and economic advantages for your organization which can be created by an industry investment in the DSDC.

I encourage and request that you submit feedback and questions to the DSDC feedback form (https://www.digitalsciencedev.net/feedback-form) so that your questions can be answered in future posts. If you prefer, reach out directly for more information.

Sincerely, John Harman
Founder, Digital Science Development Consortium
Harman Solutions LLC