Article

Dark data: How using overlooked information can prove vital in manufacturing R&D

Version 1.2

Dr. Dirk Ortloff

July 17, 2020

https://www.do-it-service.de/

This article shows the importance of and solution approach for overcoming the dark data challenge in R&D. It shows the effects, high costs and project delays caused by unleveraged data when developing fabrication technologies.

1 Introduction

In times of growing international competition and shrinking market niches, innovating via new or improved products becomes key to success or even to survival. When developing a new product or a new iteration of an existing product, it's vital to use information from a variety of sources to help shape the process and ensure success.

That said, access to historical information, knowledge and wisdom is also vital to ensure that any new project learns and moves on from those completed beforehand. Experiences gained from existing tactics, previous R&D, scientific papers and old lab books provide the major contribution to the successful realization of new manufacturing processes and products. Lessons can be learned, insights can be applied to new projects and even failures or mistakes can be vital in terms of determining how not to approach an issue in a future project.

2 Building an information-driven R&D function

For manufacturers, access to historical information, knowledge and wisdom is the crucial factor in ensuring process and quality control throughout the Research and Development (R&D) function. In diverse high-tech industries, for example semiconductors, MEMS and nanomaterials, the experiences and information gained from previous projects provide essential insight when planning and executing a new development.

The advent of Industry 4.0 has also introduced the concept of a Digital Twin: a virtual version of a machine, a process or a product, which is composed of the data that they generate. Together with a model of how the real object operates, interacts and evolves, a digital twin can be of enormous help in making informed decisions. With machines, for example, this can provide invaluable context for how machines operate at their most efficient, helping to roadmap the decisions needed to reach a certain objective.

However, too often these different types of data are not being recorded, shared or used – which has given rise to the term "dark data". Significant amounts of information are either ignored, not saved in the correct location or format, or simply not captured in the first place. Asides from being a waste of valuable time and resource, not looking at or learning from this data can have a serious impact on the efficiency of a business and in particular, its manufacturing and R&D operations.

3 The emergence of "dark data"

However, the reliability and accessibility of these sources is not ever guaranteed. Colleagues can move on and not leave complete records, meaning it's not always clear which experiments have been carried out and what the results and findings were. Lab books are a great resource for historical data, but are often only useful to those who wrote them. In terms of computer files, they can be distributed over several file servers, desktops and laptops. Every engineer has their own way of saving and storing information, so various software is used and there is often no consistency in the way data is stored. This leads to a build-up of dark data and in turn, a lack of knowledge.

Studies show [Barkai 2012] that in manufacturing R&D activities, approximately 40% of experiments are repeated. Experts in semiconductor process development estimate that 10-15% of failed and double experiments could be avoided, if previous results were more easily accessible. Furthermore, when engineers move from one project to another, there is a risk that experiments can be jeopardized as a new engineer arrives to be greeted with a flood of unstructured and unfamiliar data.

When documenting experiment data in R&D projects, the focus should always be on collaborative knowledge management, enabling multi-user access and clear search criteria. However, this can cause problems, as data can be structured (tables, numbers, units) or unstructured (images, emails, documents) and studies have found that up to 80% of a firm's digitized data is unstructured.

Even the structured data may change hourly or daily, and if old or out of date information is not deleted, archived or correctly labeled, inconsistencies and confusions can arise. Searching data can also be problematic, as search criteria and reporting can change project to project and even within one project, depending on the task in question. Added to this, full text search is often not enough to pinpoint the exact data needed. It also doesn't give the context of the data, such as how the results were achieved, where else the same material or conditions were used or how a certain component was produced.

This build-up of dark data leads to the undesirable result that certain information is only used within the live cycle of one component or project. This limits the learning for future projects, as only those who were involved in the project or enters the data themselves can either access it or understand its context.

4 Solving the dark data issue

Despite these challenges, there are strategies and processes that can be put in place by manufacturers to capitalize on dark data and realize significant efficiencies in how they work on new R&D projects. The IT and technology we have access to now allows for sophisticated and intelligent systems of recording, accessing and sorting data, so these problems can be easily overcome.

On one hand, the fact that we can store more data than ever before is a big positive. However, as the amount we can store increases by around 50% every year, there is a risk that we can drown in that data if it is not correctly categorized and recorded in a way that makes it useful for future projects.

The solution is to find a more intelligent approach to documenting and managing data, establishing a framework for collaborative learning which is geared towards R&D priorities and delivers faster, cheaper and more efficient routes to market.

In order for the data to work well for R&D, any system must be able to store both structured and unstructured data, manage the relations between all the data and give every entry and link a specific description and explicit meaning. In many ways, this makes it similar to a "Single Source of Truth" solution in that it enables you to manage all of your data in one place.

It must provide an audit trail which details what changes were made, when and why, and whether they apply to the structured or unstructured data. It must also cater for multi-disciplined working environments, e.g. providing the electrical engineer with electrical test data and the mechanical engineer with stress test data.

ne of the key considerations when finding an R&D data solution is the format. Usually, R&D data is accumulated and stored in the following ways:

- **Semi-structured data in spreadsheets, text and CSV files (commonly MS Excel)**: data can be easily imported and exported through a variety of software packages and many manufacturing machines export their data in Excel or CSV format.
- **File-based result data from diverse metrology tools**: this is for digital experiment data images, analysis results and diagrams. Context can be easily lost here as the relations between data points are complex and not compatible with search functions on many systems.
- **Existing lab databases**: Collecting data from all experiments for traceability purposes; highly valuable in terms of optimising manufacturing operations and providing R&D insight.

5 PDES: a modern solution

Although all of these formats work in their own way, there are limitations on their value to the R&D process unless they are joined up and connected so the data can be contextualised and analyzed properly. By using Process Development Execution System (PDES) [Wikipedia 2020] software, these existing data silos can be integrated and merged (in one place or virtually linked together) to create a more usable and insightful set of historical information that can be applied to future projects.

The camLine XperiDesk software [camLine GmbH 2020] is an example of PDES implementation, which streamlines R&D activities by collecting data, tracking it through its life cycle and organizing into usable activity blocks. Data from existing silos can be integrated by copying or linking into the software, delivering historical insight which can inform and improve the performance of the current project.

Tools within the software allow users to load and link data from various sources, manage it in its full context and retrieve it as actionable information, thereby providing a full circle approach to the data management process.

As technologies continue to evolve and manufacturing capabilities continue to increase, it's vital that data from historical experiments is not only collected, but used in a way which enhances future innovations without wasting time or resources. PDES is a huge step forward in overcoming the issues caused by the dark data phenomenon and providing engineers with workable, reliable data from multiple sources which can help them achieve greater efficiencies in their work.

References

Barkai, Joe (2012): "Accelerating Science-Led Innovation for Competitive Advantage". In: **url**: http://accelrys.com/resource-center/white-papers/accelerating-science-led-innovation-request.php (cit. on p. 3).

camLine GmbH (June 26, 2020): *XperiDesk – Streamlining Technology Development*. In: **url**: https://xperidesk.com (cit. on p. 4).

Wikipedia (June 26, 2020): *Process Development Execution System*. In: **url**: https://en.wikipedia.org/wiki/Process_development_execution_system (cit. on p. 4).