# Surfing the Rich Data Deluge

## STEPS TOWARD DEVELOPING AN EFFECTIVE IT STRATEGY

**By John Russell**, Contributing Editor, *Bio·IT World*

Produced by Cambridge Healthtech Media Group Custom Publishing

Tessella
Technology & Consulting

www.tessella.com

# Surfing the Rich Data Deluge

## Steps Toward Developing an Effective IT Strategy

### How Did We Get Into This Mess?

It's hardly a secret that today's pharmaceutical and biotech companies are drowning under a deluge of digital images and other types of rich data. From NextGen DNA sequencing to high content screening (HCS) to in vivo whole animal imaging, these complex and rapidly evolving technologies are opening important avenues of scientific exploration across drug discovery and development. However, they also confront IT organizations with the challenge of managing large, diverse data sets.

There are, of course, many types of rich data, and covering the specifics of how best to manage all of them is beyond the scope of a single brief paper. Instead, this paper will focus on digital imaging data, which is perhaps the most varied of all rich data and remains the most widely used in drug discovery and development. Importantly, the issues around managing digital images reflect rich data management challenges generally.

In the 17th century Anton van Leeuwenhoek, a Dutch cloth merchant and pioneer in lens grinding, peered through an early microscope and famously saw a multitude of "little beasties" in stagnant water. Since then, imaging systems have matured into powerful technologies used by scientists across the entire spectrum of pharmaceutical drug development to dissect intracellular biochemical pathways, generate biomarkers, and monitor disease progression.

Modern imaging technologies use sensitive image acquisition devices and advanced signal processing capabilities to resolve, interpret and display digital images. These systems are finely tuned for specific purposes and even subtle differences in scientific need can justify purchasing new technology platforms. As a result, the sheer number and diversity of imaging platforms required to support R&D activities continues to expand rapidly.

Digital imaging platforms have significant IT requirements, from databases and servers and disk stores to sophisticated image analysis. IT support for these systems is not confined exclusively to server rooms but now extends into the laboratory setting. Sometimes IT organizations were involved platform selection discussions to ensure technology support and integration costs were understood and planned up-front. More often than not, those important discussions never happened. The rapid maturation of imaging technologies from standalone systems to sophisticated computational platforms simply outpaced the realization that conversations were required.

Because IT was not involved in the technology platform selection, researchers did not budget for corporate SAN storage. Instead, image files were stored on unsecured disks under lab benches or office desks at risk of being lost.

Even more importantly, the challenges of incompatible technology platforms, proprietary file types, technology limitations and the lack of frameworks to integrate this information in a meaningful way with other corporate knowledge databases went unexplored. The ongoing M&A frenzy compounded these problems by bringing together large, diverse and sometimes incompatible experimental and IT platforms under the same corporate roof.

Not surprisingly drug developers are being forced to rethink their rich data management strat-



egies. Business leaders are focused on wringing full value from their existing investments while simultaneously adopting the latest technologies as they come to market. IT leaders are struggling to store and integrate the large volume of images already in the corporate environment, let alone support emerging technologies.

## View from the Imaging Trenches

"The big problem is the technology is charging ahead and scientists are pushing ahead at a pace which I think they themselves don't realize or appreciate just how much data they are going to be generating. That's one issue," says Stephen Emby, Imaging Management, Research/Business technology informatics, Regenerative Medicine, Pfizer. Once data start piling up, "they realize they've got a problem and page IT for help. These local IT groups can help manage the data but what you see is a whole disparate range of solutions, with different types of hardware being adopted, different standards and so forth. So there's a disconnect."

Emby cites a specific instance in which "scientists were doing ultrasound work and a two-week ultrasound study would take them three months to analyze. The reason it took so long was they hadn't thought about the data flows and hadn't engaged with IT in any way," says Emby. "We got involved and quickly identified ways we could cut that down significantly, saving huge amounts of time for the scientists."

More broadly, Emby says, "Another issue, particularly with large corporate structures, is that IT infrastructures move at almost a glacial pace with regard to upgrades because the costs involved are high. Our infrastructures are built to handle gigabytes of data, not terabytes of data or petabytes of data. So that's another disconnect."

"The view I've been trying to instill is we need is a global strategy for managing this kind of data, one that not only copes with the current situation but also is planning ahead for what inevitably is going to be an ever growing volume of data and increasing diversity of data as well."

To overcome the deluge of rich data, IT organizations must seek solutions in four key areas:

- **IT Infrastructure.** Most IT systems handling image and rich data have been cobbled together over time. These fragmented solutions frequently include vendor specific platforms that don't play well together. There are many internally developed solutions, which tend not to scale well. Rich data storage is often scattered and disconnected. Enabling and managing collaboration around rich data use is another thorny issue and can be difficult.
- **Conflicting Needs.** Scientists and business decision makers often have different requirements (s/w, data formats, etc). Cultural and technical issues (work practices, tools, far-flung locations) often impede data sharing throughout the enterprise. Regulatory requirements for development may prove burdensome and unnecessary for discovery scientists.
- **Standardization & Governance.** Standardizing platforms and suppliers is a worthy goal but difficult to achieve and often at odds with rapidly delivering a system to meet scientific needs. Image and data retention policies are often lacking. Workflow policies (tools, data formats, annotation requirements, visualization s/w, etc.) usually need strengthening.
- **Cost.** Decisions made in virtually any of the forgoing areas impact cost. Software (instruments, analysis s/w, image visualizers, etc.) and hardware (disk, server, workstation) costs are the most obvious. Support costs are also often higher because of the fragmented nature of most rich data management systems. Given the cost-control imperative rippling through the industry, all expenses face a higher justification threshold.

## The Tessella Approach

Generating a comprehensive IT roadmap for rich data management is a challenging and painstaking task that requires deep domain knowledge and IT expertise. To help companies assess road-

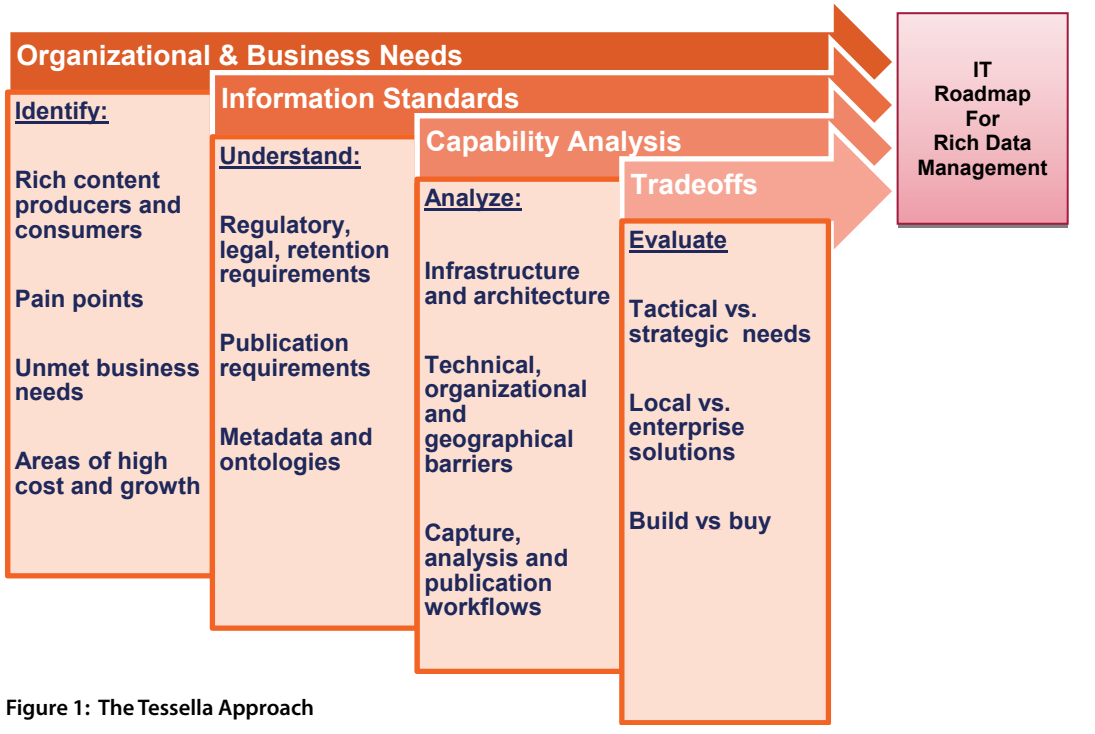


**Figure 1: The Tessella Approach**

map options, Tessella uses a proven consulting process to generate IT strategies that maximize return on investment in rich data technologies (Figure 1).

Depending upon the size of the company and the scope of the project, developing a roadmap takes roughly 1-to-4 months. Through a combination of structured interviews and careful assessment of client capabilities using their rich data management capability model (Figure 2), Tessella helps companies define an IT roadmap tailored to their specific business requirements.

The reward is worth the effort. Reducing the cost to exploit the value of rich data (from omic to image) from across functions and improving knowledge interchange helps project teams reduce the risk in study, assay and project decision-making. Effective long term scientific data management and knowledge planning offers R&D teams with sustained, faster access to rich data for exploration in a changing corporate environment.

## Identify Organizational & Business Needs

An important first step of Tessella's approach is to identify business and IT stakeholders and define organizational and business priorities. Typically, the challenge is to identify all relevant producer and consumer stakeholders who are often dispersed throughout the R&D organization, across geographies, and even across company boundaries (e.g. CROs). Once identified, the stakeholders participate in structured interviews to define the business challenges they face. Questions tackled include:

- What types of images do they generate and for what purpose?
- Approximately how much content do they generate?
- Where are the IT pain points?
- What are the high level unmet business needs?
- Which image types are most costly to manage and where is growth greatest?

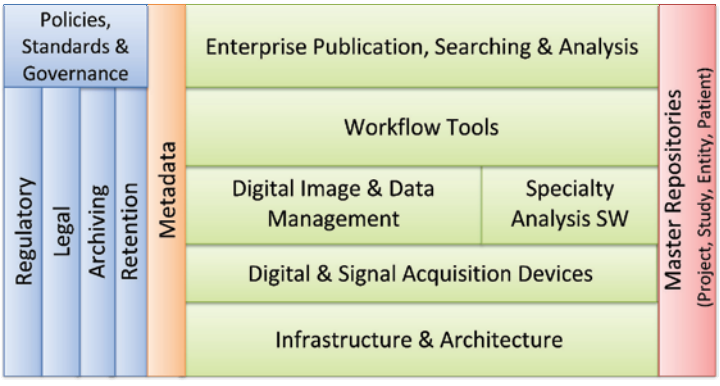


**Figure 2: Tessella's Rich Data Management Capability Model**

Next, in consultation with the IT leadership, Tessella determines specific priorities (e.g. operational costs, business and time-to value/benefits, etc.) so business priorities are understood and subsequent efforts can be focused on achieving maximal impact.

## Understand Existing Information Standards

Because rich content generating devices came into widespread use quickly and organically, necessary policies, standards and governance measures are frequently incomplete. For example, companies might lack clear image retention and archiving guidelines, or enforce these unevenly across the drug discovery pipeline.

Establishing clear standards around what data is stored, in what format, for how long, and what annotation is required should be essential pieces of a company's digital image retention policy. "Scientists' natural inclination is to keep all of their images because they don't know when they might want to go back and reanalyze them," says Dr. Scott Shepard, Senior Vice President for international consulting firm Tessella.

"At the same time, we often find instances where scientists fail to load key experimental results derived from those images into corporate repositories. For example, scientists frequently use Excel to analyze the derived data and either don't load key results into corporate repositories or do so only at the end of a study as more of a clean-up exercise," says Shepard. "Often the images and derived data are stored on unsecured drives from USB sticks to DVDs to personal laptops. Until this information is loaded into corporate repositories, it is at a high risk of loss. Our customers value our ability to quickly identify these producers and bring them into compliance."

Defining data standards, especially metadata standards, is key to ensuring knowledge gleaned from rich content can be reused by scientific researchers. Standards enable rich content information to be accessible via corporate search capabilities and facilitate integration of rich content with structured data generated in other parts of the company, leading to fully-informed decisions and faster generation of comprehensive data packages required for regulatory submissions.

## Analyze Current Capabilities and Future Trends

In the future, says Emby, managing, sharing, and reanalyzing image data will become even more important. "Just as people are mining sequence data today, I think mining image data will become more and more desirable," he says. "One researcher may be scanning rat liver sections and another is scanning rat kidneys section or human kidney section. Although they are scanning samples for their particular interests somebody else may be able to reanalyze that data for their own interest."

"As image analysis becomes more and more sophisticated, more people will want to tap into those images and reuse them discover things they weren't originally designed to show but the data is there and they can now pull that stuff out. I think that's where this is going to go and that will be the next explosion," Emby says.

The challenge of managing the large volumes of

diverse rich content types in one geographic location is difficult enough but many companies need to share these large files among collaborators at research sites and research partners around the world. Because today's networking technologies limit the ability to transmit rich content in real-time, companies must make pragmatic decisions about their hardware architecture and consider tradeoffs of storing images locally versus low-cost enterprise stores.

## Creating Customized IT Roadmap – It's all about tradeoffs

It goes without saying that having a robust software and hardware infrastructure and architecture is fundamental to managing rich content, but "we frequently find an array of gaps that significantly impede the ability of scientists to capture, store, analyze, and publish digital rich content," says Dr. John Godfree, Senior consultant at Tessella.

Creating rich content IT Roadmaps is a difficult balancing act fraught with competing business needs and priorities. The IT pain points don't always align with the needs of scientific researchers and today's reduced IT budgets make it impossible to fulfill all needs. "Companies receive tremendous benefit by having an impartial facilitator with deep rich content expertise to identify gaps and broker difficult discussion around these inevitable tradeoffs," explains Godfree.

Since costs for the IT infrastructure needed to handle rich data is already significant and growing rapidly, IT leaders often push scientists to make use with existing technology solutions. Scientists often chafe at these decisions because the tools are too rigid to meet their existing requirements and do not accommodate the latest technologies just coming to market.

To the extent practical, standardizing on components is preferable. "It reduces licensing costs, maintenance costs, and ongoing operations costs. IT has a more simplified infrastructure to manage which in turn frees time for IT to spend on more pressing areas of the business plus stay abreast of

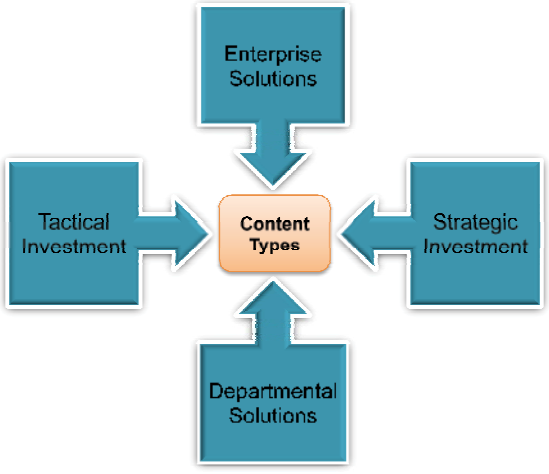


**Figure 3: Finding The Right Balance**

newer technologies such as hierarchical storage and cloud technologies that help reduce the costs," says Shepard.

Yet there is no one-size-fits-all approach, so companies have a difficult time creating enterprise-wide standards. Rather, they need to define the business needs first, then understand if existing standards are sufficient or if new standards are required.

Ultimately, the roadmap must find the proper balance between competing factors (Figure 3):

1. When should enterprise systems be used and when should content be kept locally
2. How should tactical investments to support immediate business needs be balanced against future strategic investments?

Time-to-value and cost considerations are important factors to finding the proper balance. "Because the cost and time to integrate new data types with corporate systems is high, tactical solutions sometimes provide better value propositions for certain types of rich data," says Godfree. Little value on investment is gained by physically integrating rich content data into a standard corporate repository if the content is not widely used, for example. In these cases it makes better sense to retain the files in a local repository with or without pointers from enterprise search tools. "Certainly, we en-



courage use of standard platforms where possible, but in our experience, companies need to accept that not all rich content will be integrated into enterprise stores anytime soon," says Godfree. "The cost of doing so simply outweighs the benefits."

This does not mean there aren't significant opportunities to reduce costs of managing rich data. Strengthening policies around data retention and improving compliance with those policies is a relatively easy step companies can take to reduce the storage costs of these files.

Take, for example, high content screening (HCS) systems that typically generate many small JPEG files that are relatively easily integrated into corporate repositories. When this technology is used in routine drug optimization screening studies, these images can be of high value and worthwhile capturing in enterprise systems. However, it makes less sense to capture the same files generated in high throughput screening campaigns where most images lack useful information.

## Conclusion

It is virtually impossible to effectively conduct modern drug discovery and development without relying upon advancing technologies – whether they are NextGen sequencing machines or HCS platforms or something else - and the huge volumes of digital image and rich data they produce. Moreover this dependency will only grow as new instruments are developed and introduced to help scientists at all phases of the R&D pipeline.

While the scientific and competitive imperatives will drive adoption of these sophisticated tools, the economic reality is now and will continue to force the biopharmaceutical industry to seek ways to control costs, become more efficient in their R&D activities, and derive the maximum ROI from these investments. The only way to accomplish this is to develop a clear scientific and IT roadmaps surfing this growing flood of image and rich data. Tessella is ideally suited to help companies develop and implement such a road map.

Tessella
Technology & Consulting

www.tessella.com

### TESSELLA CORPORATE DESCRIPTION

Founded in 1980, Tessella is the international provider of science powered technology and consulting services. World leading organizations choose our unique blend of science, engineering and sector expertise to deliver innovative and cost-effective solutions to complex real-world commercial and technical challenges. Our people are high achievers from leading universities and are passionate about delivering value to clients. We are proud that our work makes the world a better place to live in: developing smarter drug trials; preserving the digital heritage of nations across the globe; minimizing risk in oil and gas exploration; controlling the orbit and attitude of satellites; researching fusion energy.

**Tessella plc** 26 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire OX14 3YS, UK
T: +44 (0)1235 555511 | F: +44 (0)1235 553301 | E: info@tessella.com

**Tessella Inc** 233 Needham Street, Suite 300, Newton, MA 02464, USA
T: 1 617 454 1220 | F: 1 617 454 1001 | E: info@tessella.com

Tessella
Technology & Consulting

www.tessella.com