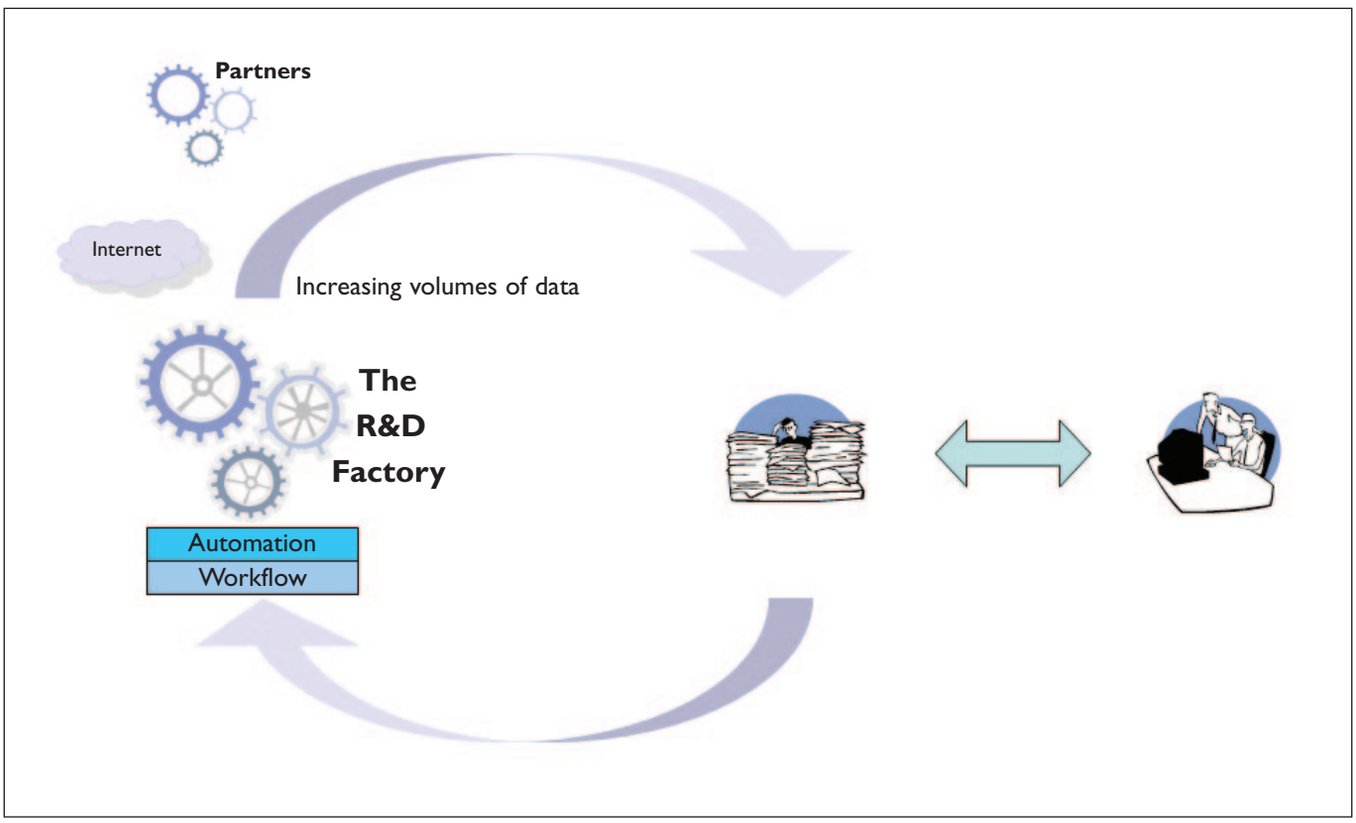
Unlocking the value in DATA

In recent years enormous sums have been spent on systems and data management, but are we getting full value from these investments? We discuss here how organisations, scientists and, indeed, vendors themselves must work as one to extract total value from our vast data repositories.

By Dr Amber Salzman

In the era of High Throughput Chemistry, High Throughput Biology and Genomics, R&D has become industrialised in many ways. To support this way of working, millions have been spent on systems and data management. The R&D Factory approach has resulted in the generation of huge volumes of data – Genomic, Proteomic, Transcriptomic, High Throughput Screening, High Throughput Biology, Safety and so on. For the most part, the data has been used in support of these specific vertical operations. We have not gained full value from these investments by looking across the data generated. While advances in computer science have improved integration and searching, many challenges remain. The lack of consistent usage of terms is rate limiting in many areas. Even terms such as project, disease state and therapeutic area have different meanings depending upon where they are used in the R&D process. Other limitations include our ability to make better use of textual information and data stored locally to particular users. In addition, we should be able to gain more insights from effectively linking proprietary and non-proprietary information. While it is early days for the general integration and accessibility problem, there are very effective solutions in place for specific areas. Also, there are pragmatic approaches that can be taken to make best use of existing capabilities and position for the future.

When tackling these issues, there is a general desire to integrate everything, thinking that will solve all and the gems will flow forth. Vendors often position themselves as able to do this. Terms such as middleware, portal and data transformation are thrown about like magic dust that can be sprinkled on the data repositories to release the gems. Given that the systems and processes have evolved in the absence of a well-defined enterprise information model and in the absence of standards, this is not a reasonable expectation. However, today's tools and capabilities can be



leveraged to yield great value. It is important to be realistic about what can and cannot be done, and implement an approach that progressively provides more and more capability.

Focus on areas where questions are being asked and focus on areas of easy opportunity

People often raise the integration challenge at an abstract level. Brainstorming which questions would be useful to answer focuses the integration problem and enables addressing the challenge in a step wise approach. For example, many pharmaceuticals have put considerable thought into predictive toxicology and have well defined questions in this area. Data from across several areas of R&D can be used to gain insights into predicting toxicity. In some cases, there are questions that currently cannot be answered. In others, the answers are derived only after expending considerable effort. Those that are currently being answered very inefficiently are usually the low hanging fruit. One might think that the real insights only come from being able to answer questions that today cannot be answered. However, the 'answerable' questions are not always systematically asked due to the tremendous resource requirements. Another approach is to select cases where there are well defined relationships between data sets and with minimal effort the information can be linked together and made accessible with an easy-to-use front end. Once pulled together, this asset should be well positioned in a way that encourages and challenges scientists to gain insights by asking questions and viewing the information in ways they had not imagined.

Tools and approaches should be applied and pursued aggressively

While the lack of standards and varying meanings attached to words can hinder the ability to link across data sources, there are opportunities to get more out of what exists today.

Visualisation tools: Vendor supplied visualisation tools, such as Spotfire, have made it relatively easy to browse huge data sets and facilitate decision making in a way that used to be impossible. As an example, this capability can be used in the selection of lead compounds by pulling together bioassay results, calculated properties, phys-chem properties, structural motifs, sequence information and target infor-

mation. Another example is visualising genetic information with assay data. While this is incredibly powerful, the limitation comes in when considering the next level of relationship such as linking the assays to pathways. It is important to note the strength of these tools in support of browsing data marts, while being aware of the difficulty in drilling down to follow relationships between data.

Flexible system design: When relationships are well understood, the opportunities should not be missed. Systems can be deployed to support these relationships. For example, the relationship between compounds and assays is well understood, and most companies have designed their systems to support viewing what assays have been applied to which compounds. When systems evolve over time without thinking about ways to make them flexible enough to accommodate new opportunities, it takes considerable effort to compensate. As an example, in many companies R&D has taken on a virtual approach, utilising partnerships and alliances. However, this approach was not baked into processes and systems until recently. As a result, when considering intellectual property rights, considerable effort is needed to amend systems to support viewing compounds that are free of alliance implications. There is no simple way to link alliance data and compound databases.

Portal technology: In cases where data relationships are well understood, there still exists a challenge to finding all the data. The information may be contained in various sources with a disparate collection of interfaces. In these situations a single web front end to the various sources can be employed with relative ease. Of course, this approach requires a central view to the many operational systems deployed throughout the organisation. Several companies have developed portals into their portfolio management information, supporting transparency to study information in a way

that used to require multiple phone calls. Many are now browsing their portfolio and drilling down to areas of interest. In many cases, overlapping systems had evolved over time so considerable effort was required to determine where data would be entered and maintained so that there was only one version of the truth.

Standards: While many companies can now browse their portfolio, all tell their war stories of how difficult it was to sort through what was meant by various terms. Once there were agreed pick lists for fields such as disease area and project codes, the various operational systems could then be linked easily to provide the portal view. Developing standards may not be glamorous, but it is a lynch pin to moving forward. We have all experienced the challenges of developing standards. On the surface, definitions are arbitrary so should be easily derived. However, at another level there are deep reasons and history attached to the various definitions used. Reaching agreement is not trivial. To avoid analysis paralysis, an optimal approach is to time box the decision and agree to an 80% solution, if necessary. Those held accountable for developing standards should feel comfortable escalating at times of impasse to get an 'executive decision' and move forward. There are many terms requiring standardisation. It is most helpful to have a team that understands what needs to be addressed as a priority in order to resolve the most critical integration problems first.

Ontologies: Ontologies specify how to represent objects, concepts and their attributes, as well as specifying the relationships between them. There are several vendors that have emerged recently that are providing ontologies for specific domains enabling searching across information where different terms are used. This is incredibly powerful in addressing the different meanings and different significances that people apply to the same thing. The limitation with this approach is that developing ontologies is a considerable effort and will not solve the problem of linking across data sets. Standards are critical in addressing the integration problem.

Intelligent integration of information: Several vendors have recognised the need to support searching across huge sources of non-proprietary information, without necessarily bringing it 'in-house'. Kliesli from GeneticXchange, SRS from Lion, and DiscoveryLink from IBM enable linking various proprietary information and non-proprietary information such as PubMed (literature), Swissprot and Refseq (proteins), KEGG (pathways and pathway classes), Mesh (disease/phenotype). By propagating use of these tools, the effort required to search all these sources in support of an activity can be dramatically reduced.

Centralising access and using search engines: When a dispersed community of scientists is interested in a particular domain of information, then it may be worth investing in a systematic approach to making all the information accessible. This may be as basic as allocating shared disk space and actively encouraging the community members to place their private files in the shared space, or as sophisticated as creating repositories with tagged information, ontologies and employing a search engine such as Verity. It is constructive to make use of collaborative tools such as Groove or Lotus Notes to share information between users. These tools support ad-hoc sharing, as well as planned initiatives. In some cases, assigning an information curator may be needed to progress the effort.

There are other aspects to keep in mind when approaching the way forward.

Data security can not be ignored

As delivery of integrated information becomes more and more common, information about company assets becomes more and more transparent. This is a wonderful enabler in an environment that requires so much interdisciplinary knowledge and expertise. However, the flip side is that critical company assets may be put at greater risk of leaking. Information should be categorised at a conceptual level based on the level of access that should be granted. The ability to grant access easily based on roles and categories may be more or less difficult depending on how flexibly the system was developed. It is important for an organisation to decide consciously how it wants to balance the need for transparency and engagement of its people with the need to maintain assets securely. This is not a trivial decision. It has implications on whether people feel trusted and whether access to information is viewed as hierarchical. It influences the culture of the organisation. In addition, the role-based access security model should be implemented with a standard approach, so that users do not have a different username and password for each system.

Work with source data

Since data stored in one system to serve one purpose is not always easily manipulated to serve another, data is often copied from one place to another. When working through what information should be linked, it is important to find where it originates. In some cases, it may be beneficial to create an approach to ensure data is only entered once in a way that it can be reapplied easily. This is where techniques, such as using XML and defining source systems and reporting systems, can be applied effectively. It is important to implement a strategy that stores data once, while feeding multiple views.

Search tools should be employed along with improved data tagging

We all search the Internet to find information readily, and would like to be able to search our intranets in the same way. However, it is important to be aware of how difficult it is to search effectively without knowing the semantics of what we are looking for, ie the meaning and significance is needed. In order to harness the company's information in a way that would be most meaningful to an organisation, the information must be centrally accessible, indexed, and in many cases tagged with metadata to enable searching. Until vendors make tagging more automated, this can be a considerable effort. This can be done in a step wise approach, addressing most critical data first, and making use of powerful searching tools and ontologies.

Vendors may have a role

In addition to vendors which can supply tools to support activities such as portal development, searching, ontology development and visualisation, there are vendors which consult in this general space. An organisation should decide how it wants to tackle gaining more insights from its vast repositories of data. In addition to a general approach, an organisation should define the operational steps that will move it towards reaping the true value of its investments.

Conclusion

An organisation must examine its particular business challenges and determine where it is most beneficial to work through gaining broader insights from current investments in systems and data management. To generate the untapped value, capable technology specialists should make best use of the newer techniques and tools available, and facilitate the organisations review of opportunities. Scientists throughout the organisation should be engaged and pressed to generate insightful questions that will drive us to unlock the value from our vast data sources. DDW

Dr Amber Salzman *currently serves as Senior Vice-President of R&D IT at GlaxoSmithKline. She has been with GSK and legacy companies for more than 20 years. Prior to that she worked at Medical College of Pennsylvania. She holds a PhD in Mathematics from Bryn Mawr College and a BA in Computer Science from Temple University.*