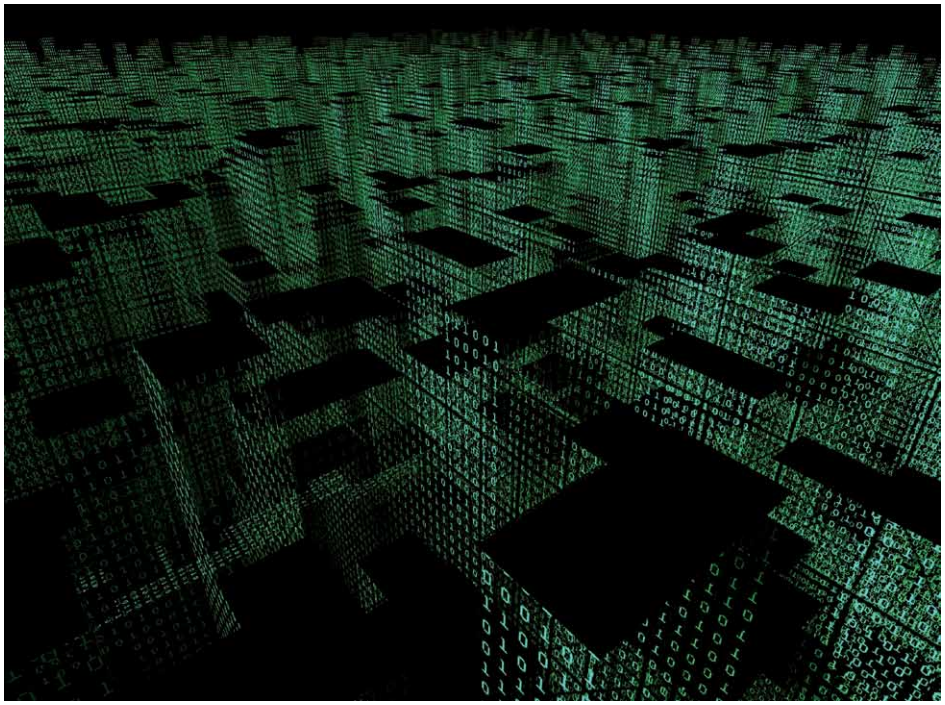

Data archiving in a digital world

Road map to archiving structured data

Jeroen Tegelaar, Peter Kuiters and Jan Geurtsen

Data archiving is often unjustly seen as a troublesome and time-consuming activity that needs to be performed after a process or task has been completed; a necessary evil that has to be done at least once a year. In this article, we argue that data archiving can deliver strategic added value and therefore deserves more attention and appreciation. The misunderstandings around this theme are explained and a definition of the term is followed by an overview of the benefits. The article further covers the different types of solutions that are available for the implementation of data archiving, with attention for specific criteria in the choice of application or tooling.

J.A.C. Tegelaar
is a senior manager at KPMG Management Consulting, IT Advisory.
tegelaar.jeroen@kpmg.nl

P. Kuiters
is a manager at KPMG Management Consulting, IT Advisory.
kuiters.peter@kpmg.nl

J.M.B. Geurtsen
is a consultant at KPMG Management Consulting, IT Advisory.
geurtsen.jan@kpmg.nl

Introduction

Within the discipline of data management, archiving (referred to as 'data archiving' in this article) holds a special position. It is often unjustly regarded as the stage that heralds the end of a process, whereas archiving is ideally the domain *par excellence* where structured data (read: data stored and ordered in databases) and unstructured data (read: e-mail, documents, social media content, etc.) complement one another in such a way that, in an ideal situation, Enterprise Data Management can be achieved: the situation in which all information is available and accessible to the right person at the right time for the right purpose.

This however is generally not the case in real-life, because overdue maintenance, especially with regard to born-digital data, precludes such a situation. Remarkably, most organizations are aware of this problem but take no action, even if they know that this will eventually lead to unac-

ceptable risks and high maintenance costs in the future. The cause of this is easy to guess: responsible managers do not know how to approach archiving because traditional methodologies for the administration of physical archives are no longer applicable.[1]

Accordingly, IT management organizations tend to see data archiving mainly as an interesting selection instrument to distinguish between data that is immediately useful to everyday operations and other kinds of data. The first group must be constantly available and accessible. The rest can be stored somewhere else, offline or in the cloud, at much less expense, until the agreed retention limits have expired.

People are insufficiently aware of the fact that digital data requires a different approach, one that not only solves today's problems but also provides new opportunities. To be clear about what data archiving can yield in a digital world, we shall first elaborate on what we mean by this concept and further examine the misunderstandings that have arisen around this topic.

Some misunderstandings about data archiving

There are lots of misunderstandings about data archiving. The most common one is that storage is the same as archiving. This is incorrect. Simple storage means an unmonitored back-up of data, with various methods and storage media being used. Archiving goes much further. The value of data is determined, relevant information is selected, storage limits are defined, and measures to protect the data are determined based on the location of the data and the anticipated need for (long-term) access. After all, archiving means that data is still actively managed, on the precondition that the data itself can no longer change. The difference between these two methods is the structured approach and use of policy measures. Data archiving is the second to last stage in the lifecycle of data, the next and final stage is purging information.

Archiving is not the same as storage

A second major misunderstanding is that activities involving data archiving only need to be initiated at the end of the process. Particularly with regard to unstructured data, the earlier you know which data will be created and how and for what purpose it will be retrieved once stored, the sooner you can estimate the value of the data. The results of this are that the data can be better protected and, at the same time, distributed to the right people over its lifecycle. This approach also makes clear at an earlier time which retention times ought to be applied, so that the IT management organization is aided in its goal to keep the cost of data management as low as possible.

A third misunderstanding is the equation of data archiving with the process of making back-ups. Having a back-up is a necessary business continuity measure, and is directed toward recovery after a system failure. In fact, attention is paid to the availability of an IT application by means of which a business process is supported after any calamity. In the case of archiving, however, it is the data that is most important, while the software is of secondary importance. Data often survives the application in which it is stored or, as mentioned earlier, it is migrated to an alternative environment in order to create space. This may be another, cheaper location or a new application. Back-ups of data may be a means to archive data, but in real life this is generally not the primary reason for making back-ups.

In our view, data archiving is not only the archiving of digital data but also the result of the whole body of processes, procedures, activities, measures and resources that aim at storing data in such an unchangeable manner that, for a fixed period, accessibility and availability can be realized for the right person at the right time. In this article, we direct our attention to the archiving of born-digital data in the IT domain, although we do realize that the archiving of physical documents can be seen as a component of data archiving. Furthermore, the focus of this article is on the challenges of archiving structured data. Unstructured data is essentially different in terms of the trigger and the approach, and therefore requires a different kind of methodology.

1. An article entitled 'De tijd van aanmodderen is voorbij' (The Time of Muddling Has Gone, see [Glas04]), dated 2004, argues that organizations are insufficiently aware that the criteria that apply to physical documents and corresponding management measures are no longer suited in a digital world, because the characteristics of digital data differ essentially from the characteristics of data recorded on paper. The fact that the distribution of data currently runs via more informal channels such as e-mail, web-based applications and internet, instead of the previous, formal (and still existent) channels such as post rooms and secretariats, demands a (radically) different approach to the process of archiving data.

Use of and need for data archiving

Now that it has become evident what data archiving means, we can now examine the added value that data archiving can deliver.

Data archiving to save costs on server space has just been mentioned as a reason to undertake such an effort, but there are many more reasons that may put data archiving high on the agenda of management. The implementation of place- and time-independent employment is a first important reason, where the desire to work independently of place and time can only be given substance if information is available and accessible in the right form at any given moment. For this, the digital archives must be in good order; otherwise personnel will not be able to find what they are seeking. Another frequently mentioned reason is the phase-out of applications, where the organization wants to or must retain the data stored in those applications, for example due to contractual obligations. Data archiving must meet legal obligations concerning the retention of information, as specified by the Dutch Tax Department for example, but also the timely destruction of personal data on the basis of the Privacy Act often provides a reason to instigate data archiving. Data archiving is occasionally placed high on the agenda because an organization is confronted with the high cost of a lost legal case because a certain item of evidence could no longer be traced. Finally worth mentioning: data archiving is better for the environment. This is because less hardware is needed to manage the lower amount of data that survives beyond its expiry date. This is translated into a lower use of energy and therefore a lower strain on the environment.

Organizations and the value of data

The above-listed motives can be clustered into four different groups of arguments, as graphically shown and explained in Figure 1.[2]

In this context, it should be immediately mentioned that most organizations are not aware that the reasons given often serve different interests. And these interests may be opposing ones.

For example, adhering to the principle of accountability, an organization may wish to satisfy certain criteria based on rules and regulations. But, in doing so, it may ignore the impact that certain regulations may have on everyday business operations, and vice versa. The use and management of personal data is a good example. From the perspective of the conduct of business, the personal data of customers may provide valuable insight for marketing and sales goals. The personal data of one's own staff, received from several sources such as staff satisfaction surveys or job performance interviews for example, may contain information about the current business culture and the quality of business operations. From the perspective of responsibility, however, this data should strictly

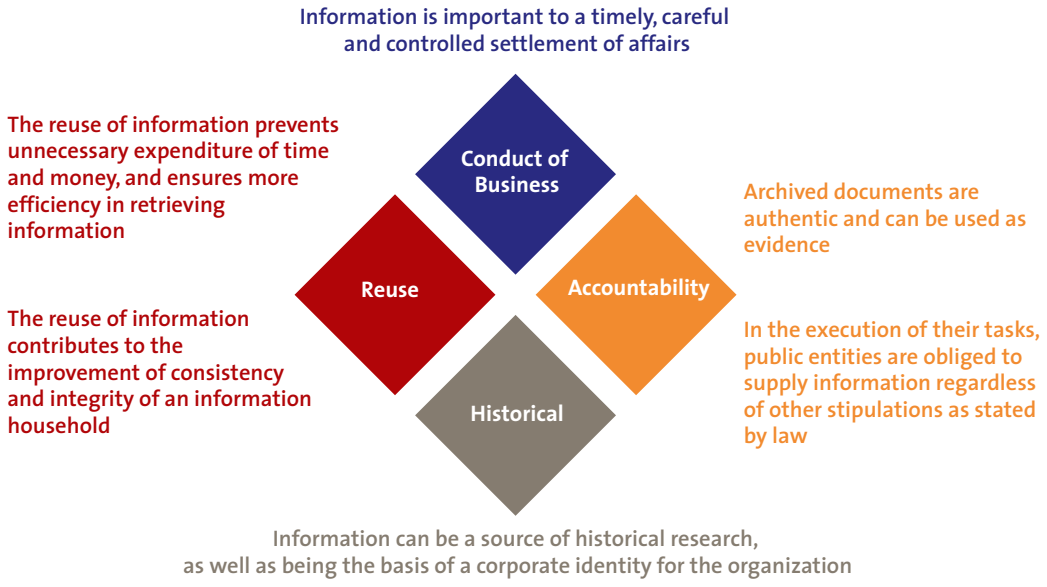

Figure 1. **Value of archives.**

2. The article by Eric Ketelaar, professor emeritus in Archival Sciences at the University of Amsterdam, entitled 'De waarde(n) van archieven' (The Value of Archives), mentions the above-stated importance of archiving data. Moreover, he insists that many people unjustly see business operators as 'creature without memory', after an analogy drawn by Marc Boch. The archives are seen as being a glory hole for old information that is not relevant to the reality of the competitive marketplace. However, Ketelaar observes that large organizations with a long history of commercial success, such as Coca Cola and Shell for example, actually do recognize the value of historical archives, in the form of income from the reuse of old advertising material and as a basis for the defense of patents and brands. Shell carefully keeps the reports on all oil drilling activities since 1890. They can refer back to them if a long-abandoned well suddenly gets a new lease of life thanks to new exploration techniques. See also: [Kete04].

only be used for the objectives for which it was collected in the first place, and after use it ought to be immediately and irretrievably destroyed. If the department responsible for this data does not realize which effect may arise from storing these data illicitly – without complying with legal stipulations – the organization will run the risk of being fined by the designated supervisory authority (in this case of privacy in the Netherlands, the CBP). Even more important than the fine (which is rather modest for most organizations) is the potential damage to the company reputation. This is often a multiple of the initial fine. Setting up and embedding a good archiving strategy, and incorporating the various demands that an organization makes on information, is certainly no superfluous luxury.

Scenarios for data archiving

When data is not available, it has a negative impact on everyday business operations, as the organization incurs costs that are not covered by turnover. The magnitude of these costs depends on the impact of the lack of this specific data or, in other words, the business value of the data. Accordingly, the first step in data archiving is always to make an analysis of the value of the data. This shows the strategic value of data archiving. The temporary and occasionally permanent absence of this data always forms an unacceptable risk that must be mitigated by a combination of solutions, measures and procedures. For this reason, the above-mentioned analysis is commonly a strategic risk analysis of the data. A binary glance at the data, where only the issue of whether or not the data ought to be kept, is much too limited. This will lead to conflicts in the deliberation of interests. Without counting the costs of storage and management, this may lead to arbitrariness and high management costs because people opt to save everything 'just in case'. By weighing risks as a factor (on a scale from 1 to 10 for example) a decision can be made between risks and costs of migration. This gives a more qualified and balanced conclusion that forms the basis of an archiving strategy, which is the second necessary step within the domain of data archiving following the risk analysis. Of course it may still be absolutely necessary to keep data, in cases of legal obligation for example, but this will always be more of an exception than the rule and such data should then also be assigned an exceptional status.

Judging the data and its value, as well as its use and the data protection required, is an important step in the preparation of data archiving. With this insight, we can now take a step toward implementation. Before dealing with the aspects relevant to the choice of a specific solution, we shall describe a number of different scenarios.

The market offers a great many methodologies and solutions for data archiving. To cut a path through this surfeit of choices, we structure these solutions across two axes. First of all, there is the object of archiving, the application or the data. Second, there is the issue of centralization versus decentralization.

In an application-oriented strategy, existing applications are in fact left intact. In the simplest form, this may mean the continuation of the existing server. When the object of archiving concerns data, the relevant data is selected and stored in a different form outside the application. Decentralized solutions have a tactical significance and are based on several mutually independent datasets, opting for a separate archiving solution for each application, for example. With centralization, the data is stored in one single central solution, better known by the definition 'data warehousing'.

When these characteristics are set against one another, a matrix of four basic archiving strategies is created (see Figure 2).[3] These are:

1. continuation: in which existing applications are continued independently of one another in largely unchanged form
2. virtualization: in which existing applications are converted to virtual servers and share a pool of available hardware
3. tooling: where the relevant data is stored in an application-specific tool
4. managed: where the data is stored and managed in a central tool.

In this context it is important to note that this matrix is not a fixed set of four variants but rather a continuum for the various solutions. The four scenarios are described in the remainder of this section, outlining the advantages and disadvantages, but presenting only the broad contours.

3. Based on a combination of existing methodologies of BiSL, ITIL and ISO-15489, as well as experiences in real life.

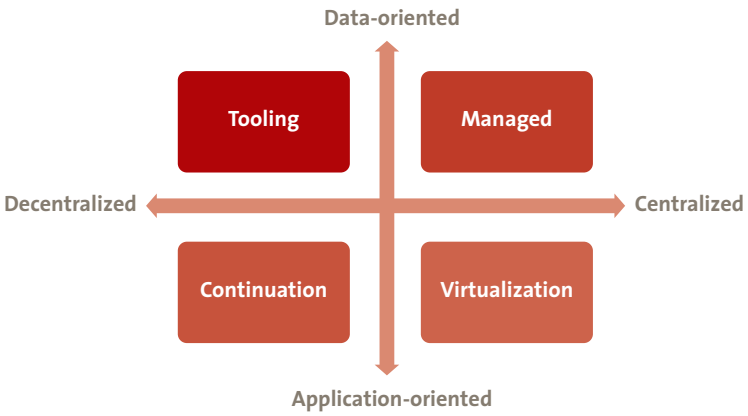

Figure 2. **Four basic archiving strategies.**

Scenario 1. Continuation

The continuation scenario is the most simple and also the most frequently applied. The implementation is straightforward: in fact, nothing is done. Existing applications with the underlying server simply remain in place. In our experience, the primary reason for this strategy being applied is the lack of a methodology for archiving. As soon as a legacy application has to be decommissioned after the implementation of a new one, there is always the doubt as to whether or not all the relevant data, such as historical data for example, has been properly transferred or otherwise recorded. In order to avoid any risk, people tend to leave the old applications intact, next to the new application suite.

However, keeping the existing servers going has only disadvantages. First of all there are the technical management costs: the existing environment has often been set up on a large scale, which is generally unnecessary from an archiving perspective. With a cost structure based on capacity, it is not difficult to incur large costs. At the same time, these dormant systems may form an obstruction to data consolidation, where a large amount of legacy servers can push up costs.

Another disadvantage is the fact that hardware is subject to obsolescence; or perhaps the configuration of the system may have to be adjusted due to certification and maintenance requirements specified by the supplier. Approximately six years is the maximum lifespan of an item of hardware; for data, this is more or less a minimum. As a result, it is often difficult to implement the necessary system improvements in the fields of hardware, the operating system and application software. For smaller applications, especially when these are stored on a virtualized server, the impact is smaller than it is for larger applications such as ERP, CRM, HRM and SCM systems. In the case of this last group of software, application expertise will have to be kept up to date by the company's own management organization.

Scenario 2. Virtualization

Virtualization is a good way to reduce the costs of management of little-used applications. The core of the solution is that archiving applications can be deactivated for most of the time. As a consequence, it is only necessary to store the so-called 'image', and management costs are reduced by switching off the environments. Both the storage and the pool of virtual hardware are centrally regulated, meaning that this strategy is fundamentally different from the continuation strategy.

In this scenario, the necessary hardware is found in the available overcapacity in one's own data center or in the cloud. In fact, the cloud can be an excellent option for larger applications; a large chunk of extra capacity is needed only temporarily. Of course, there are extra preconditions that need to be met when dealing with the cloud, such as the location of data storage for example. However, such choices can be determined for each instance and local servers can be deployed as an alternative.

By keeping the virtualization environment up-to-date, the dependence on old hardware is removed and flexibility in the data center is increased. In the light of a fixed data retention period, the chosen configuration may extend beyond the maintenance contract of the supplier, while an upgrade may be necessary. Therefore, it is clear that this scenario is only suitable when access to the archived applications is incidental. Similar to the previous scenario, knowledge of the archived applications must be maintained. With storage times that may extend to ten years, this may be a troublesome task and may be contrary to rationalization initiatives for the rest of the application landscape.

Scenario 3. Tooling

With the deployment of tooling, the emphasis is on the relevant data, with it being transferred to application-

Continuation of the existing situation is the most expensive and risky option

With the Managed scenario, learning a single archiving solution is sufficient to gain access to all relevant data

specific tools. It can be expected that this tool will retain the application context (look-and-feel and data structure), thus enhancing user acceptation. For a single or small-scale solution, this is a good alternative, and is often the selected option in the absence of a broad archiving strategy. An important aspect of this scenario is that a subset of data is selected. After all, only the data in the tools is kept, the rest is deleted. This data selection is a useful, but no trivial occurrence.

A simple approach to tooling is the storage of archive data in the form of reports. Juridical reporting obligations, such as those imposed by the Dutch Tax Department, form the starting points in this context. The obligatory reports are formulated and stored in a durable form, such as PDF-A. This is sufficient to meet the storage criteria imposed by the Dutch Tax Department (see also). It is evident that no ad-hoc access to the data is possible with this strategy. The future usefulness of data stored in this way is rather low.

With regard to application-oriented solutions, the management costs are often significantly lower. This is due to the fact that most of the data is deleted, so that storage costs decrease. Another advantage is with the supporting hardware and software. By opting for a separate solution and by keeping it up to date, the data remains in a certified configuration supported by the supplier. The ability to make use of modern (virtual) hardware is an extra advantage that contributes positively to the continuity and flexibility of the landscape.

Tooling is not ideally suited to a broader approach because the data is distributed across several archiving tools, so that relationships between data are difficult to identify. In addition, knowledge of various archiving tools must be kept up to date.

Scenario 4. Managed

When the data is centralized in a single solution, we have created a managed strategy for data archiving. In this scenario, the data management strategy implements the entire data lifecycle and is observed with the aid of the chosen archiving solution. The implementation of a central solution is often the most expensive option. This is because the solution seldom provides resources to archive all the applications in question. As a result, the effort required to execute a controlled data transfer is often greater.

1. All data is placed in the same system and only one archiving solution needs to be learned to gain access to all relevant data.
2. It becomes possible to establish relationships between data from different systems without having to make use of system interfaces.

Particularly the second point can be advantageous in unlocking the value of the data, by connecting to a data warehouse for example. In this way, data can have a greater function in terms of reuse, or in discovering trends by applying data-mining techniques. Because all data is managed centrally, maximum profit can be gained from advanced solutions in which data policies are automatically set in motion. In this context, we can think of the automatic removal of sensitive data after the expiration of the retention period. This will help reduce the cost of management activity.

Relevant criteria in the choice of a (software) solution

After determining the value of data and its anticipated use, account should be taken of the following aspects in the formulation of an archiving strategy and the choice of solution:

- user acceptation
- security and continuity
- functionality
- costs.

User acceptation

The non-acceptation of an archiving solution by the users is devastating for the successful completion of an archiving process. Old applications are not cleared away,

or historical data is not cleansed or transferred from the existing application. In neither case will the intended benefits be obtained. Without any attempt to be exhaustive here, we name the following aspects that lead to a more ready acceptance of the strategy and solution by users:

- Retention of application context. The look-and-feel and data structure of the archived data is similar to that of the source application. This aspect is primarily relevant to a tactical archiving solution (tooling).
- User friendliness. A solution that is pleasant to use is essential to user acceptance, especially when there is a change of application context. The solution is made cost-efficient by the lack of necessity to intensively train personnel to work with the software.

Security and continuity

As indicated previously, the stored data has value to the organization. A reliable solution, in which access to data is in line with competencies within the organization, is necessary to do full justice to the value of the data. This is especially a factor when data from several systems is combined. This not only concerns security against external access, in cases involving personal data for example, but also against internal access where staff members from various legal bodies may not have access to one another's financial data. Continuity can be guaranteed by means of common back-up procedures.

Functionality

Each application provides different functionalities and, correspondingly, also archiving solutions. Ad-hoc request procedures are a good example here. These must harmonize with the organization's everyday practice and there must be equilibrium between the effort expended and the result obtained with regard to the request. Matters such as supporting documents to structured data, such as scanned suppliers' invoices, clearly have added value.

Costs

Almost every conceivable combination of user acceptation, security and functionality can be realized. However, a higher score on the various components pushes the cost of investment up. In the process of coming to a fitting solution, the before mentioned risk analysis can again be applied.

Example: archiving ERP system

An organization replaces the supplier of its ERP system. At the implementation, the decision is made to transfer only the current data; the historical data is not transferred. However, this data ought to be stored, in connection with legal obligations. There is no archiving strategy, but the topic is high on the agenda.

The preservation of the application environment is regarded as undesirable, due to license, hardware and management costs. Virtualization is impossible due to the frequent access requirements for stored data. There has been no choice in favor of a central tool, so a tactical solution for this ERP environment (tooling scenario) seems to be the most appropriate option. The cost of implementation of this tool is significantly lower than the preservation of the existing system.

The most convincing arguments in favor of this tool are:

- the costs of implementation, where the data migration tools supplied are of great importance;
- the user friendliness, where the preservation of application context is very important.

With regard to security and functionality, there appeared to be little distinguishing capacity between the solutions reviewed. The organization in question eventually opted for the tooling scenario, because this seemed to offer the best balance for good embedding in the existing management organization in relation to the relatively low costs of migration.

Conclusions

The sections provided above have led to the following conclusions:

- Data archiving should not be seen as a problem or as a technical trick of the IT domain, but should rather be approached strategically in order to visualize the true added value provided.

Data archiving should be approached as a strategic advantage if added value is to materialize

- Although the reasons for data archiving may often differ, the principles behind data archiving do not. These can always be traced back to the need for an efficient conduct of business, accountability for the work performed, reuse of information, and historic goals.
- Data archiving can be viewed from the perspective of the application or data. However, a prior analysis of the value of data is advisable in order to keep the costs of the project under control.
- Four strategic scenarios can be discerned with regard to data archiving. The scenario of Continuation can be seen as the most expensive and hazardous. The Virtualization scenario is the most interesting from a management or technical point of view. The Tooling scenario is the most practical solution. Finally, there is also the Managed scenario, which would appear to be the most valuable in the long run, but entails the greatest initial investment.
- When choosing the solution, account must be taken of user acceptation, functionality, security, business continuity and, of course, costs.

References

[Bela10] Belastingdienst 2010, *Uw geautomatiseerde administratie en de fiscale bewaarplicht.*

[Glas04] Boudien J. Glashouwer RE RI CISA and Jan Pasmooij RA RE RO, *De tijd van aanmodderen is voorbij*, Controllers Magazine 2004/3.

[Kete04] Prof. Dr F.C.J. Ketelaar, *De waarde(n) van archieven*, Archievenblad 108/2, March 2004, pp. 16-19.

About the authors

J.A.C. Tegelaar is a senior manager at KPMG Management Consulting, IT Advisory. He has obtained a great deal of experience with the strategic, tactical and operational aspects of archiving procedures, and functions as the point of address within the Service Line Enterprise Data Management in connection with queries in the field of Document Management and Records Management.

P. Kuiters is a manager at KPMG Management Consulting, IT Advisory. He has had much experience with the configuration of various types of architecture (from Enterprise to IT) and the supervision of processes in which architecture is examined. His clients are mainly involved in challenges based on the control of complex heterogeneous application landscapes.

J.M.B. Geurtsen is a consultant at KPMG Management Consulting, IT Advisory. In the past few years he has been concerned with information security and data archiving, particularly in the financial sector.