

A Service-Oriented Architecture for a Flexible Digital Public Archive

Henk Jonkers, Christian Wartena, Hugo ter Doest

This paper reports on the development of a service-oriented architecture in a project that aims to realise a digital archive for long time preservation of digital objects of the municipality of Rotterdam¹. The architecture is based on the Records Management Continuum, which argues that archiving starts with the creation of an object. The infrastructure, system architecture, archiving processes and services are specified according to this principle. The paper concludes with a discussion of the advantages of service orientation for this project, and based on this, recommendations for similar projects with an infrastructural component undertaken by local governments.

Introduction

Governmental organisations have a legal obligation to archive their records in an ordered and accessible condition. Both national and local governments have organised their paper archives in order to meet these legal obligations. However, at the moment there is a shift from analogue (paper) records to digital records, and a growing number of objects to be archived is "digital born". Also, at the moment of creation of these objects it is not always clear what contextual (meta)data must be captured in order to facilitate archiving at a later stage. An additional problem with digital objects is that there is a risk that they can no longer be read and interpreted in the future, due to changing digital formats. This development requires a careful interpretation of the Archiefwet, the Dutch law article that regulates archiving for governments and public institutions. For instance, the Archiefwet requires that archives must be ordered and accessible, but the meaning of these terms in a digital context is not yet established. Besides the handling of digital born objects, processes and systems must handle records consisting of both analogue and digital objects and "digital reborn" (digitised) objects.

The city of Rotterdam is currently reorganising its information and technology policy by adopting an enterprise-wide service-oriented architecture. While to some extent information policies currently are still bound to single departments, in the near future shared services will replace department specific solutions. One of the reasons for this approach is that more and more processes are executed across multiple departments instead of within a single department. Another reason is that governments are required by the national government to offer their services to citizens and companies online (one of the goals of the program "Andere Overheid" of the Dutch Government). As a consequence, it is more effective and efficient to provide digital archiving services that are used enterprise-wide. Rotterdam has decided to develop a service-oriented architecture as a blueprint for the ICT facilities.

¹ The work reported on in this paper is the result of a project on behalf of the Gemeente Archief Rotterdam (GAR) in 2006. This paper reflects the opinions of the authors that are not necessarily shared by the GAR. The text and the models presented in this paper are based on material created in the project, but some minor improvements have been made. The GAR has approved publication of this article.

One of the main characteristics of the service-oriented architecture (SOA) paradigm is that it separates services from their realisation. For Rotterdam, this offers the possibility to integrate existing systems and new services in the same architecture. Existing systems are hidden behind service interfaces and new systems are developed from the start with the architecture in mind.

The core business of the Gemeente Archief Rotterdam (Municipal Archives of Rotterdam, GAR) is permanent and reliable storage of archives and collections according to legislation and regulations. Furthermore, departments and a wide cross-section of the public must be able to access this information. GAR regularly inspects the archiving practice of the departments and advises departments on their information policy and the organisation of their archives. Early in the development of a digital archive, GAR found that the service of long-term digital preservation of municipal records is an enterprise-wide issue that needs to be addressed on a higher organisational level [Horsman and Pompe, 2004].

This paper is organised as follows. In the next section we discuss long-term digital preservation in more detail; in particular, we focus on the principle of the Records Management Continuum. In Section 3 we outline the service-oriented architecture we proposed for the digital public archive. Furthermore, we show how the architecture can be positioned in the municipal architecture. Section 4 discusses the advantages of a service-oriented approach in general. Section 5 concludes this paper with a discussion of the advantages of SOA for this project, and based on this, we give recommendations for similar projects with a strong infrastructural component undertaken by local governments.

Long-term digital preservation

Long-term preservation comprises all activities that are necessary to guarantee long term accessibility of data. In many cases, records have to be archived not only for historical interest but primarily to document administrative and decision-making processes. Archived records eventually can help to solve conflicts. A scaring example is provided by the investment bank Stanley Morgan. Recently, the bank was unable to provide evidence relating to a claim against them by a former client. The relevant correspondence was done by e-mail. The e-mails from the early 1990s could either not be found or not be read. A Florida court found that the bank should settle the \$603 million claim, plus \$850 million punitive damages².

For digital documents, long-term preservation does not only mean that data have to be stored conscientiously, but also that arrangements have to be made to ensure the accessibility and readability of the data. This problem is specific for digital data, since bits and bytes cannot be read directly but need the mediation of specific hardware and software. Since both hardware for reading storage media and software for rendering data change constantly, this is not a trivial problem. Furthermore, identity, integrity and completeness, authenticity and (contextual) understandability of the files have to be ensured. This problem is not unique to digital files; nevertheless, it requires a somewhat different approach as compared to the traditional practice, since the digital record keeping process differs from the traditional one. Moreover, problems of multiple versions and manipulation require other solutions in the digital world than in the world of paper files. Of course, there are numerous dependencies between these aspects of preservation. According to [Caplan, 2005], the quality aspects of a preservation system can be ordered hierarchically, resulting in the Preservation Pyramid (Figure 1). The pyramid shows the main problems for digital preservation, whereas the words to the right indicate the strategies to handle these problems. According to this model, moreover, a problem can only be handled if the problems lower in the hierarchy are solved.

² <http://www.tessella.com/news/CostOfDPFailure.htm>

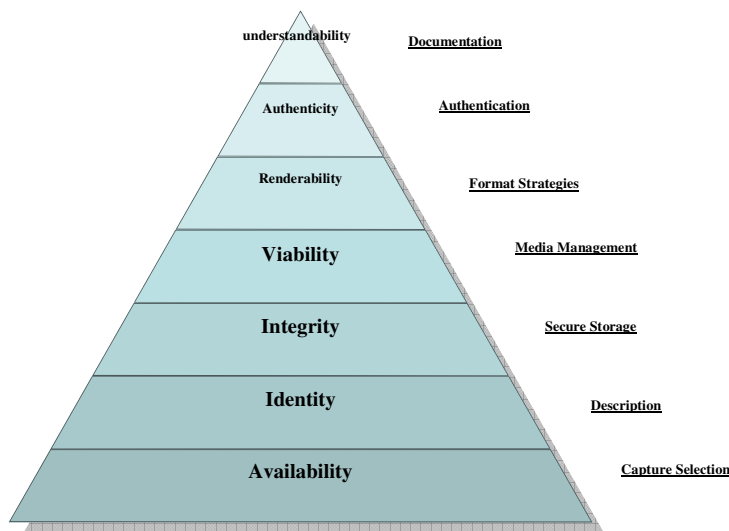


Figure 1 Preservation Pyramid after [Caplan, 2005].

The first type of problems, those concerning the systems and file formats becoming out of date, is either solved by emulation techniques or by migration and conversion (usually in addition to preservation of the original format). In the first case, the original software that is needed for rendering the data is archived along with the data. A virtual machine is used to emulate the original environment in order to run the software in the future. Now only the virtual machine has to be kept up to date with actual hardware and operating systems. In the second approach, files are periodically migrated to new versions of the used format or converted to more sustainable formats. In fact, there are many preservation strategies in between, and actual systems might use hybrid approaches. We refer to [Thibodeau, 2002] for an overview.

The second set of problems for preservation of digital data, those concerning integrity, completeness, identity and authenticity is closely related to the way documents are created and managed. While it might be problematic for traditional paper records to reconstruct missing information on the creation process etc., for the vast amount of electronic records produced today this is almost impossible. Thus all the information that is needed to enable successful archiving needs already to be present as from the creation of each document. This principle is one of the central aspects of the *Records Management Continuum* [Upward, 1996].

In traditional archival science there is a strong focus on the life cycle of a document. According to the life cycle model, the life of a document can be divided in several phases. A prevailing division is that into the creation, the dynamic phase, the semi-static phase and the static phase. These phases often correspond to different depositories. In the dynamic phase the document is, e.g., stored in the office of the agency that created it. In the semi-static phase, the document is no longer actively used but still important for that agency. Now it is, e.g., stored in a central depository of that organization. Finally, the document has only historical value and is moved to the depository of a central archive.

In the continuum model, in contrast, record keeping is defined as a *level* (in the original literature often called 'axis') of the total document management process, not as a phase. Similarly, actual usage is also considered a level (namely the transactional level), instead of a phase. Thus, it follows naturally that record keeping issues already can (and should) be considered while the document is still in use. The records continuum model has turned out to be a fruitful model when talking about electronic record keeping. In the first place because we no longer have a clear marking of the boundaries between phases by documents physically moving to different locations. More important is the fact that many of the issues in digital preservation have to be taken up early in the life time of a document. However, this has huge consequences for the organization of all processes concerning document management and archiving. It is not very likely that departments will hand over their documents to the archival department as long as these documents still are relevant for the department. On the other hand, the problems concerning archiving during the semi-static phase is felt as a burden.

Thus, offering high quality (digital) preservation might be a solution from which all parties can benefit: the departments are released from the problems of (mid-term) preservation without losing control and responsibility. For the archival department it offers a possibility to capture all relevant information necessary for long term preservation as early as possible. Thus the quality of all records that are finally delivered to the archive can be ensured to be good enough for preservation.

For the reasons sketched above we believe that SOA offers an optimal solution for digital preservation in large organizations. In the sequel we will investigate how this can be realized in practice.

Service-oriented architecture of a digital public archive

In this section, we describe an enterprise architecture which provides guidelines for the design of a digital archive (called the 'E-depot') for the city of Rotterdam. The architecture covers the business, application and technology layers, as well as their relations: e.g., how the archiving (business) process is supported by applications, and which technological components are used to run the applications. The service concept plays a central role in linking, in particular, the application layer and the business layer; i.e., application functionality is exposed as services, which are used in the business process. For expressing the enterprise architecture models, we make use of the ArchiMate language [Jonkers et al., 2004]. Figure 2 gives an overview of the ArchiMate concepts and relations used in the models in this article.

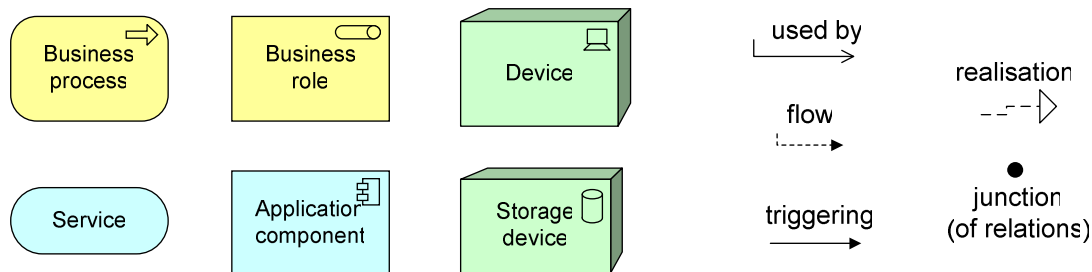


Figure 2 ArchiMate Symbols

Archiving process

As a starting point for describing the architecture of a digital archive, we take the transactional level (or axis) of the records continuum model. Thus we focus on the activities undertaken in the conduct of affairs as shown in Figure 3. Speaking in terms of the life cycle model, the first steps in the process, object creation, ordering and closing, are part of the dynamic phase of an object. In this phase, object may still be accessed or modified regularly. Therefore, they are not yet stored in the E-depot, but still kept in their original application (e.g., a Document Management System or a process-specific application) and in the original data format. The services for the creation, ordering and closing are therefore application-specific, and their realisation is outside the scope of the E-depot architecture. In the 'semi-static phase', directly after the closing of an object, it is already transferred to the E-depot; this in contrast to most of the related developments, where only objects in the static phase are stored in an E-depot. Our approach follows naturally from the continuum model and the abandoning of the prevalence of the life cycle phases, and allows us to take preservation measures at an earlier stage than would be possible otherwise.

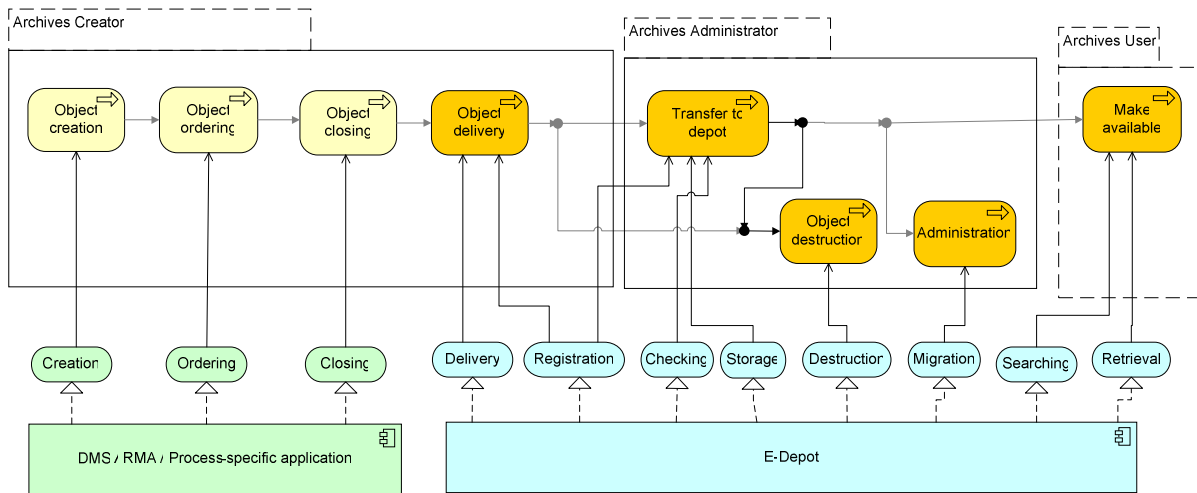


Figure 3 Archiving process with supporting automated services

Because objects are already stored in the semi-static phase, the E-depot also has to provide facilities for the destruction of objects which do not need to be stored permanently.

Components of the E-Depot

For the identification of the main components of the E-depot, the "Reference Model for an Open Archival Information System" (OAIS) has been followed [CCSDS, 2002], which is widely accepted in this domain. For the central system components, Ingest and Archival Storage, we have identified more detailed subcomponents, as shown in Figure 4.

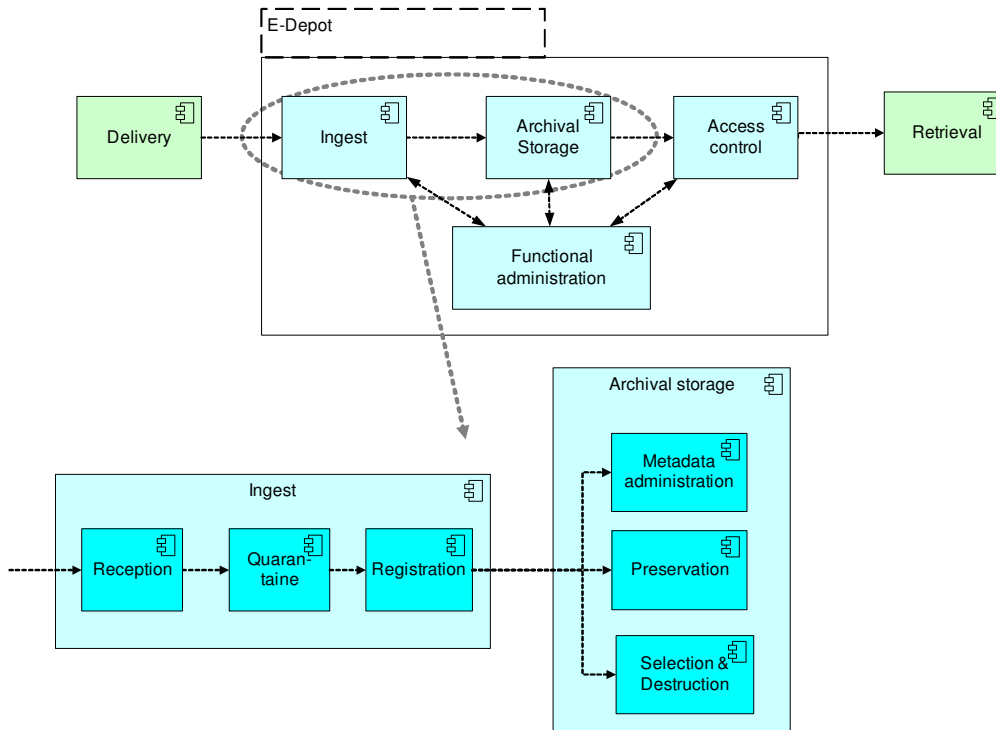


Figure 4 Components of the E-Depot

In this figure, Ingest is divided into three components. Quarantine takes care of checking viruses, including a period of isolated storage and a second check with updated virus definitions. Registration covers the important aspect of assigning the correct metadata to the records. The Metadata administration component within the archival component refers to the management of the actual metadata and the indexing of the records and files. The

Preservation component constitutes the core component for guaranteeing long term availability by applying the preferred preservation strategy.

Realisation of services

Figure 5 shows which of the services offered by the E-depot are realised by which components. It also shows which roles in the archiving process (archives creator, administrator or user) make use of which services. Note that the data flows between application components could also be realised by (internal) application-to-application services. We have chosen not to explicitly model this, thus leaving more freedom for the implementation: the application designer can decide to use services to realise the flows, but may also opt for another solution.

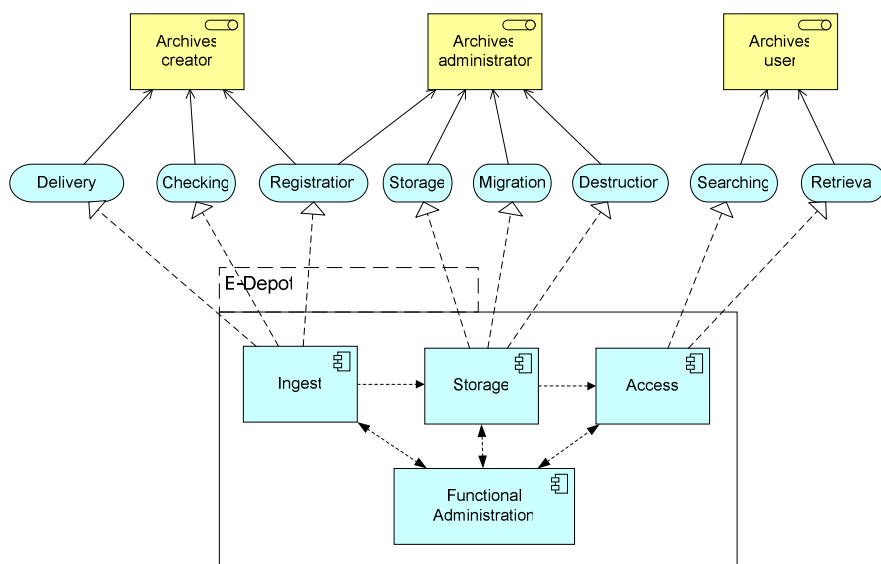


Figure 5 Realisation of services by E-Depot components

Each E-depot service should be realised in such a way that it is not aware of the other services, and thus could be realised by an independent component or a third party. An example of a solution following this principle is the so-called Dark Archive of the Florida Center for Library Automation, that offers a storage service, including preservation, but does not provide facilities for ingest and access [Caplan, 2005].

Another important application of service orientation is to connect document management systems or process-specific application used by the municipal departments to the E-depot. In this approach the applications are considered services that offer access to documents and data managed by departments. See also [Meijer and Van den Broek, 2007] for an explanation of this idea.

Mapping of components to hardware

Finally, Figure 6 shows which of the E-depot components are running on which hardware servers. There are separate servers for quarantine, registration, depot and retrieval. Furthermore, there is a single SAN device for centralised storage of all archived objects.

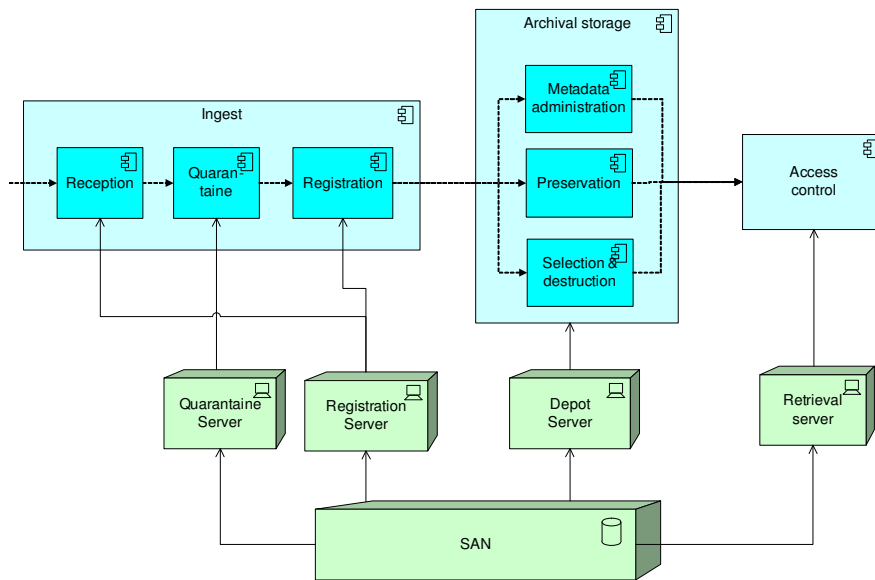


Figure 6 Mapping of components to hardware

Position of the E-depot within the municipal architecture

In the (service-oriented) ICT reference architecture of Rotterdam, the Enterprise Service Bus (ESB), on which applications can offer their services as Web services, has a central position. This also applies to the archiving services as offered by the E-depot. The services identified in the previous section can serve as a basis for this. Archiving in general plays a role in different places, both in the primary business processes and the supporting processes. The specific archiving services as offered by the E-depot could be generic services in the reference architecture, similar to document management services (see Figure 7).

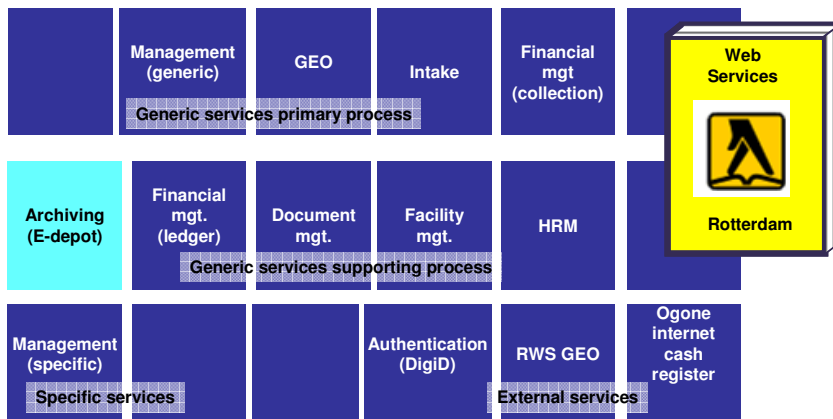


Figure 7 Position of the E-depot within the municipal service-oriented reference architecture

Advantages of service orientation

In this section, we summarise the main advantages of a service-oriented approach as identified in literature, with the emphasis of the problems that a service-oriented way of working can solve in the case of the E-depot.

Communication

For the E-depot, communication is a central issue, because many municipal departments are involved. Communication is needed with both the (potential) users of the E-depot in these departments, and the ICT departments that are responsible for connecting their applications to the E-depot. The service concept is used and understood by the different domains within the

organisation. The concept offers a common language to people from business and from ICT, which facilitates their communication. Also for the communication in a design chain, the service concept can provide more clarity.

Interoperability and flexibility

A wide variety of applications used in the municipal departments will have to be connected to the E-depot. Moreover, this application landscape is highly susceptible to change. Therefore, *interoperability* of applications is of particular importance in this situation. Related to this, *flexibility* is an important requirement: it is undesirable to be tied to specific services or service suppliers.

Service orientation has a positive effect on issues such as interoperability, flexibility, cost-effectiveness and innovative power. At the technology level, this has already gained a wide acceptance: Web services and the accompanying XML-based open standards are heralded for delivering true interoperability in this area [Stevens, 2002]. However, service orientation also promotes interoperability at higher semantic levels by minimising shared understanding between a service provider and a service user. Therefore, services may be used by parties different from the ones originally perceived. The possibility to collaborate with different partners offers new opportunities for innovation. Existing services can be combined in new ways, resulting in new or better products or services.

Interoperability and the separation of internal and external behavior offer new dimensions of flexibility:

- Flexibility to replace or substitute services in cases of failure
- Flexibility to upgrade or change services without affecting the organisation's operations
- Flexibility to change suppliers of services, thus preventing vendor lock-in
- Flexibility to reuse existing services for the provision of new products or services
- New opportunities for outsourcing

Reuse and efficiency

A common E-depot will prevent the different departments from having to develop their own solutions for the same problems of archiving. Although reuse is not an exclusive advantage of service orientation, by focussing on services, many opportunities for reuse of functionality will arise. This results in a more efficient use of existing resources. In addition, outsourcing and competition between service providers will also result in a reduction of cost.

Summary and conclusions

In local and national governmental departments, like in many administrative organisations, more and more of the traditional paper documents are being replaced by digital information objects. This also has important consequences for the archiving of these objects. In the case of digital objects, the Records Continuum model appears to be more suitable than the traditional lifecycle model that is generally applied for the archiving of paper documents.

For the city of Rotterdam, the Municipal Archives are developing a shared digital archive, the E-depot, for the archiving of digital objects produced by the municipal departments. In line with the Records Continuum philosophy, the objects are already transferred from their original applications to the E-depot at an early stage, at the start of the semi-static phase. This requires a tight integration of the E-depot with Document Management Systems and process-specific applications used by the municipal departments.

Because of the distributed nature and the heterogeneity of the application landscape in the organisation, it is important to take into account the interoperability of the E-depot with these applications throughout the design trajectory. A service-oriented approach can contribute to this. As a starting point, we have proposed a service-oriented architecture for the E-depot,

which serves as a top-level design and as a means of communication between the different stakeholders.

Additional advantages of using a service-oriented architecture, in the context of an E-depot for the city of Rotterdam, include:

- It is in line with the enterprise-wide policy. The service-oriented architecture of the E-depot seamlessly fits in the municipal architecture. The archiving services can be exposed as services on the Enterprise Service Bus, thus facilitating the interoperability with other services.
- It facilitates outsourcing. Precise service descriptions and Service Level Agreements will support cooperation with external parties and service management.
- It prevents vendor lock-in. The city or the municipal archives are in control of the architecture and are in a position to switch vendors.

Dr.ir. Henk Jonkers

Telematica Instituut

Henk.Jonkers@telin.nl

Dr. Christian Wartena

Telematica Instituut

Christian.Wartena@telin.nl

Dr.ir Hugo ter Doest

Telematica Instituut

Hugo.terDoest@telin.nl

References

- [CCSDS, 2002] CCSDS (Consultative Committee for Space Data Systems): *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 Blue Book, Jan. 2002.
- [Caplan, 2005] Caplan, P., "DAITSS: Dark Archive in the Sunshine State", presentation to the *DCC Workshop on Long-term Curation within Digital Repositories*, Cambridge England, 2005. (<http://www.fcla.edu/digitalArchive/presents/DAITSSDCC.ppt>)
- [Horsman and Pompe, 2004] Horsman, P., Pompe, K., Building a Digital Archive: A Dutch Experience, in: *RLG DigiNews*, Volume 9, Number 6, 2004. (http://www.rlg.org/en/page.php?Page_ID=20865#article2)
- [Jonkers et al., 2004] Jonkers, H., Lankhorst, M., Buuren, R. van, Hoppenbrouwers, S., Bonsangue, M. and Torre, L. van der, "Concepts for Modelling Enterprise Architectures", *International Journal of Cooperative Information Systems*, vol. 13, no. 3, Sept. 2004, pp. 257-287.
- [Meijer and Van den Broek, 2007] Meijer, A., and Van den Broek, T., *Spanning tussen ministeries en Nationaal Archief*, Automatiseringsgids nr. 36, 2007, in Dutch.

- [Upward, 1996] Upward, F., "Structuring the Records Continuum – Part One: Postcustodial principles and properties", in *Archives and Manuscripts*, vol. 24, no. 2, 1996, pp 268-285.
- [Thibodeau, 2002] Thibodeau, K., "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years", in *The State of Digital Preservation: An International Perspective*, Washington, D.C., 2002.
- [Stevens, 2002] Stevens, M., "Service-Oriented Architecture Introduction", Part 1, in <http://www.developer.com/design/article.php/1010451>, April 2002.