The Historical Data Warehouse

By Frans Smit - February 2002

Published on http://www.mybestdocs.com/smit-f-hist-wareh.htm

Frans Smit is Head of the Section of Archival Descriptions and Cataloguing at the Municipal Archives of Amsterdam. He is (and has been) also engaged in various national and international projects concerning providing access to metadata about archives and collections through search engines on the web. He reports on adapting concepts from Information and Knowledge Management (IKM) and Information and Communication Technology (ICT) into the field of organizing and giving access to metadata about historical archives and collections [1].

This paper was originally published in *Cultivate Interactive*, issue 6, 11 February 2002, and is reprinted here with the kind permission of the author and publisher.

- -

In recent decades a dramatic change has occurred in the possibilities to provide access to archives, libraries and museums.

This change is partly technologically driven and partly driven by the demands of society to provide for a greater transparency and accountability of governments. In various traditional fields of cultural heritage much effort is being made to provide integrated and instant access to information about their holdings.

This article proposes the creation of a generic model for storing data and delivering information. In this model concepts like metadata and Data Warehousing are used in an integrated way. All these seemingly new concepts are really just a new perspective on what archivists and librarians have been doing for a long time. This new perspective is necessary however in order for cultural heritage professionals to cope with demands and challenges presented by ICT [2]. The technological implementation of this model requires a sound vision on which ICT-architecture is necessary. In this ICT-architecture some important issues must be defined, e.g. which databases must be used, how content should be managed and which role standards play.

The broad objective is to create well organised content of high quality in order to fulfil the needs of the public and for an organisation to enhance its surplus value in society. These issues must create the foundation for a truly effective and integrated service to the public. However the true foundation lies within the knowledge and the consciousness that is present in an organisation. A lot of the notions and concepts in this article still need to be worked through, though as a starting point it is hopefully a contribution to the improvement of quality and quantity of services from cultural heritage institutes to the public.

An Information and Knowledge Management Model

A key issue for institutes like archives, museums and libraries is to define and to create the surplus value of a cultural heritage institute in society. This surplus value often consists primarily of giving accurate and accessible information to the public about the subjects that are the core business of the institute. This information is based upon the material that is kept in the institute, in combination with the data and knowledge about that material. In every organisation Information and Knowledge Management (IKM) is a main issue for managers. In cultural heritage institutes a good IKM is essential.

The scope of IKM includes data, information and knowledge. IKM is not a goal in itself, it is merely a necessary subject in managing an institute in order to produce accurate services to the public. As for other important issues in an organisation, like financial management and staffing policy, it is important for

organisations to start off IKM with a model-like approach. This IKM-model may differ in each organisation. A useful model that benefits from notions from other branches may look like this.

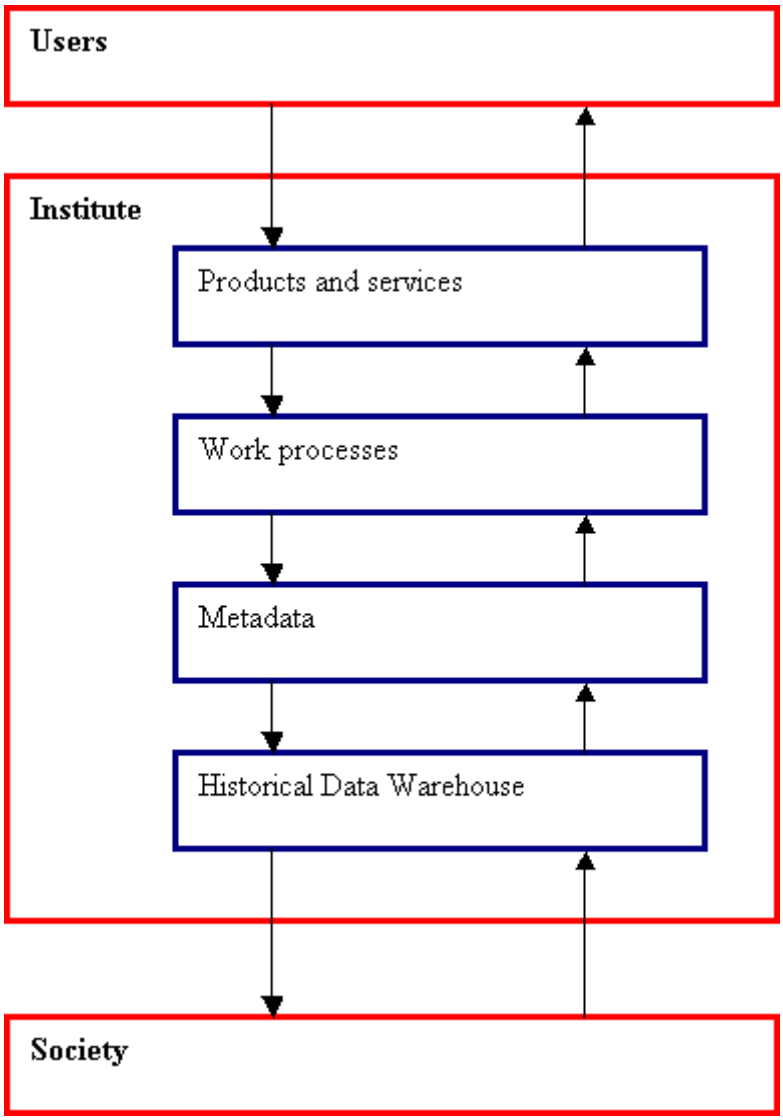

Diagram 1

The basic notion of this model is that the surplus value of the organisation consists of extracting material out of society, filtering it and enhancing it with additional information. The result of this is that the users (whoever they may be, that is not upon the organisation to determine) get better information about the subjects that the organisation is concerned with. Due to the demand of the public, services may be added, changed or removed. Every layer in this model is connected with others and is able to influence them in a direct or indirect manner.

The model has three main entities: users, organisation and society. All entities are dynamic. The organisation should take care that it does not stand between society and the users as a barrier. Its reason to exist lies in the precondition that it must give back more to the users than that it extracts from society. To realise this, every defined layer in the organisation should be designed and executed in the context of an IKM-model. Let us take a look at the organisation layers depicted above (especially focusing on the layers of Metadata and the Historical Data Warehouse).

First and foremost the products and services of the organisation show its surplus value to the public. These products may differ considerably and may include publications, exhibitions, catalogues, merchandising, and giving access to original archival material (or reproductions of that material) in a reading room. Of course a Web site should be counted among these products, creating a digital platform for a variety of services to the public. Adding, changing and removing products and services will always cause changes in the other layers in the organisation.

All the products and services are created and maintained by the work processes in the organisation. These work processes cover the whole range of activities that give added value to products created by the organisation. This range of activities is always inter-connected and inter-related to each other and to the other layers. Traditional work processes in institutes for cultural heritage are appraisal, management, description and services to the public. If you add a new service to the public, for example creating a virtual exhibit on the Internet, it will definitely have an impact on service to the public and the metadata needed to support this function. However, the process of management and cataloguing may be influenced as well. Even the appraisal of material may be affected if there is not enough original material present to create the new service.

Creating products and services to the users always require an accurate set of metadata. There is much confusion about the definition of metadata among various disciplines. In my view the concept of metadata comprises structured data about the original material that is kept in the institute (e.g. surveys, catalogues, inventories and genealogical indexes), unstructured data about that material (e.g. publications, guidelines etc.) as well as the knowledge the people in the organisation have about that material [3]. This last category may be called mobile metadata, with all the risks mobility entails (e.g. sick leaves, retirements, vacations etc.).

The accuracy of the metadata is determined by the way in which staff prepare the information that support a particular work process. The more metadata is structured and standardised, and the more they are kept in well-structured and connected databases, the better they can support the various work processes. In order to create and maintain a sound set of metadata it is necessary to have an active policy of IKM. IKM should provide for the availability for as much data as possible in a way that benefit all work processes that may need the data.

It is important to make a distinction between metadata that can be structured and metadata that can not be structured. In order to make this clear the following cycle of knowledge is a useful tool [4].

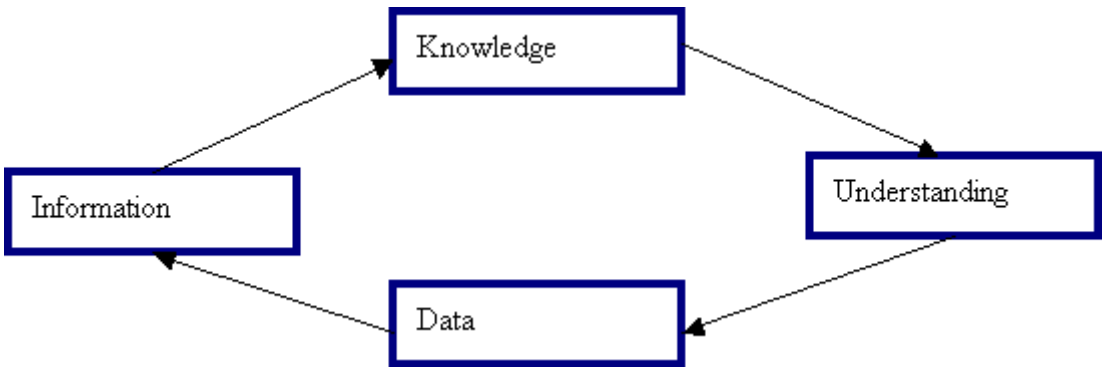

Diagram 2

In this diagram the data can be viewed as metadata in the most structured way. The data are considered "facts" and are described as such. Modern ICT-concepts and technologies are most helpful to maintain the data in such a way that it can be used in every way possible. Information may be viewed as the result of a combination of data for a specific purpose. Accurate information enhances knowledge. This knowledge improves the understanding of a particular subject. This understanding results in more data or can lead to changes in existing data.

In practice it is impossible to empty everyone's head about every subject in order to make a complete dataset of every relevant subject. The reason is that the human mind can connect and combine data and information in a way that no information system is presently capable of. It is a challenge for IKM to lay down as much information and data that is present in structured forms, so that they are easily transferable and independent of any human intervention.

For these structured forms of metadata it is possible to lay down criteria in such a way that a healthy ICT-architecture can be realised. For every organisation those criteria may differ but the following ones may be universal:

- The metadata should be up-to-date
- The metadata should be structured according to internationally accepted standards
- The metadata should always be connected to each other and to identifications of the material in the Historical Data Warehouse

- The metadata should be correct
- The metadata should be kept in state-of-the-art (that is relational) databases
- The metadata should be secure and authorized
- The metadata should be relevant to all layers in the organisation
- The metadata must be created and maintained in one work process must be available for the benefit of all work processes
- There should never be redundant metadata

Metadata can be divided into separate categories, depending on their meaning and the work process in which they are created and maintained. Possible categories for cultural heritage organisations include:

- Metadata about content
- Metadata about context
- Metadata about logistics
- Metadata about usage
- Metadata about preservation
- Metadata about appraisal
- Metadata about the physical material
- Metadata about legal aspects
- Metadata about financial aspects and insurances
- Metadata about the information systems [5].

Several accepted international standards, like Dublin Core, ISAD(G), ISAAR(CPF) and EAD, cover many or most of these categories. Not all of them however link these categories to aggregate levels and context of the metadata.

For institutes in the cultural field having the task of appraising and keeping historical material, all the products, work processes and metadata have to relate to that material. Preserving this material is on the long term the most important layer in the IKM-model presented above. In order to represent this material into a contemporary IKM-model, I have labeled it as a Historical Data Warehouse. The reason is that in ICT literature the concept of metadata is often linked to Data Warehousing. The archives and collections that should be kept in an organisation responsible for cultural heritage operate in much the same way as a Data Warehouse does in a business enterprise. It is the basis for providing accurate and inalterable information in order to facilitate decision-making and accumulate and enhance knowledge and understanding.

Marco provides the following useful definition of a Data Warehouse: "*A data warehouse is a single, entreprise-wide collection of data*". This collection should fulfil the following four preconditions:

1. A Data Warehouse is subject-oriented;
2. A Data Warehouse provides an integrated view of an enterprise's major subject areas;
3. A Data Warehouse is non-volatile;
4. A Data Warehouse holds historical views of data [6].

The concept of Data Warehousing was developed in the ICT-world for data that is kept on digital platforms. While there is a big difference with the majority of data that are kept in cultural heritage organisations this difference nevertheless does not make the comparison worthless. The preconditions of a Data Warehouse mentioned above could also prevail for archives and collections preserved on non-digital media, thus bridging a gap in concepts used in two often too separated worlds.

Another advantage to labelling archives and collections as an Historical Data Warehouse for society is that the concept of metadata falls into place. The word metadata is widely used for data that is necessary to maintain a Data Warehouse. Metadata is the data that is created, changed and removed by yourself, the data in a Data Warehouse should never be changed. As an expert on metadata and data warehousing, David Marco describes the concepts of metadata and Data Warehousing in a way that is very similar to that applied to archives and libraries: "*Metadata is the card catalog in a data warehouse. By defining the contents of a data warehouse, meta data helps users locate relevant information for analysis. In addition, meta data enables users to trace data from the data warehouse to its operational source (i.e. drill-down) and to related data in other subject areas (i.e., drill-across). By managing the structure of the data over a broad spectrum of time, it provides a context for interpreting the meaning of information*" [7].

The comparison of traditional archives and collections with modern Data Warehousing is an interesting way of putting an IKM-model for cultural heritage in the context of the digital age. Archivists were not

considered when this concept was developed. This is both surprising and regrettable because ICT-experts could have learned a lot from them about such concepts as authenticity, reliability, readability and the context and creation of data!

Information and Knowledge Management and ICT-architecture

The above described IKM-model is meant to give an overall, broad and consistent perspective on how to handle data and information in an organisation that manages historical data and material. It is like the design of a nervous system of an organisation. The functioning of this nervous system is nowadays determined by an appropriate use of ICT-systems.

In the last decades some general shifts have occurred in the usage of ICT-tools. In almost all fields, public or private, using ICT started off bottom-up, by enthusiast specialists. This phase had a character of experimenting, making mistakes and learning. With the growth of importance of ICT and with the ever-growing possibilities ICT became in a lot of fields a matter of strategic importance with a tendency to design and implement big monolithic systems. With the rise of client-server systems and especially with the rise of the Internet various systems are being used that are interconnected through a corporate concept. This concept is commonly called an ICT-architecture. The pace in which this process has been taking place varies a lot. In the field of cultural heritage institutes the described phases occur in a lot of institutes at the same time.

What is an ICT-architecture? Applegate describes it as follows: "*Just as the blueprint of a building's architecture indicates not only the structure's design but how everything –from plumbing and heating systems to the flow of traffic within the building- fits and works together, the blueprint of a firm's information architecture defines the technical computing, information management, and communications platform. The IT Architecture provides an overall picture of the range of technical options available to a firm, and, as such, it also implies the range of business options. Decisions made in building the technical IT architecture must be closely linked to decisions made in designing the IT organisation that will manage the architecture, which, in turn, must be linked to the strategy and organisation design of the firm itself. Conversely, the organisation strategy, structure, incentives, and processes strongly influence how the technology will be designed, deployed, and used within a firm*" [8].

When using the IKM-model described above, the consequence is that the usage of ICT in every layer should be accurately designed, implemented and connected to each other. At every layer ICT will play a role. In the layer of products and services one can think of Internet applications, e-commerce and applications to follow the behavior of the public. The work processes may need one or more applications to enter, change and remove metadata. The metadata it self, if digitally kept, will involve the use of databases and text-files. The Historical Data Warehouse may have digital material and digital reproductions (or even substitutes) of original, non-digital material.

Well-designed ICT-architecture must give an accurate answer to the question of how to use the various components of ICT in every layer of the IKM-model. There are various ways in making distinctions between various ICT-components. One of them is to present ICT-tools in a hierarchical order represented below.

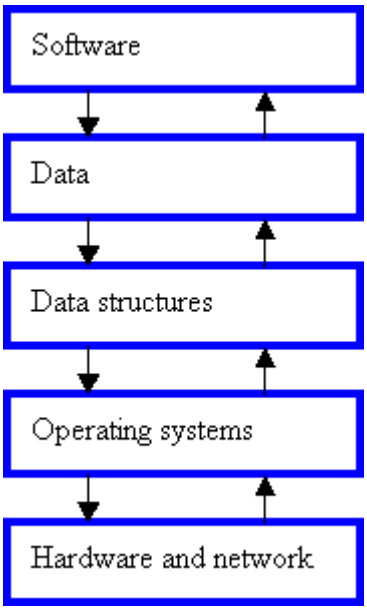

Diagram 3

As in the IKM-model every part is interconnected. On every level a choice has to be made for using the appropriate ICT products. The choices may differ according to scope, budget and structure of the organisation.

One of the ways in which it is possible to make those choices, is to make an enterprise-wide information audit and develop an ICT-policy based on the audit. The result of such an audit varies in every organisation. I will describe some conclusions that I have drawn from my daily practice.

It is very important to use state-of-the-art tools regarding data structures. The metadata layer of the IKM-model is the key layer for being able to generate useful information. Connectivity of the data is perhaps the most essential precondition. This precondition can be met by using relational databases. They are the most modern and powerful tools for designing, implementing and securing good data structures and the data itself. The language for manipulating data in a relational database, SQL, is globally accepted and supported. Relational databases are also open, which means that they can always be connected to each other in real time. Of course the logical and technical data structures in the databases should be designed and implemented in a professional way.

Standardization of metadata is essential but it is also a misunderstood topic. To those that will handle international standards regularly it may sound like nothing new, but a lot of institutes have not implemented them yet. One of the odd things about it is that those standards like ISAD(G) are often considered as being totally new, whereas they are mostly an improved version of previous standards. It is not necessary to create a totally new set of metadata, often a conversion from one data structure to another (with eventually a conversion to modern, digital media) is the only thing to be done. Another misunderstanding about standards is about when they become important. Standardization is needed to create a common language structure among organisations to enable the exchange of information. It is not important whether the database you use is entirely structured according to a standard. As long as you are able to generate your data in the desired structure when needed, there is no need to worry. Much more important is that you choose a state-of-the-art database platform so that you are assured that you can anticipate as good as possible on future developments in databases and metadata standards.

One of the fundamental principles in the ICT-architecture is that maintaining the metadata is different from its presentation. Metadata should never be redundant and metadata that belongs together should be stored and maintained together in preferably one database. You can present them however in various ways. You can present a catalogue on the Internet through a search engine, but you can use the same metadata in a printed version and you can deliver them to a Web site where they are merged with metadata from other institutes. A further important notion is that software for presenting metadata tend to change more often than software for maintaining metadata. An ICT-architecture for maintaining and presenting metadata, that takes these aspects into account and can act as a frame of reference for developing good and efficient software, may look like this.

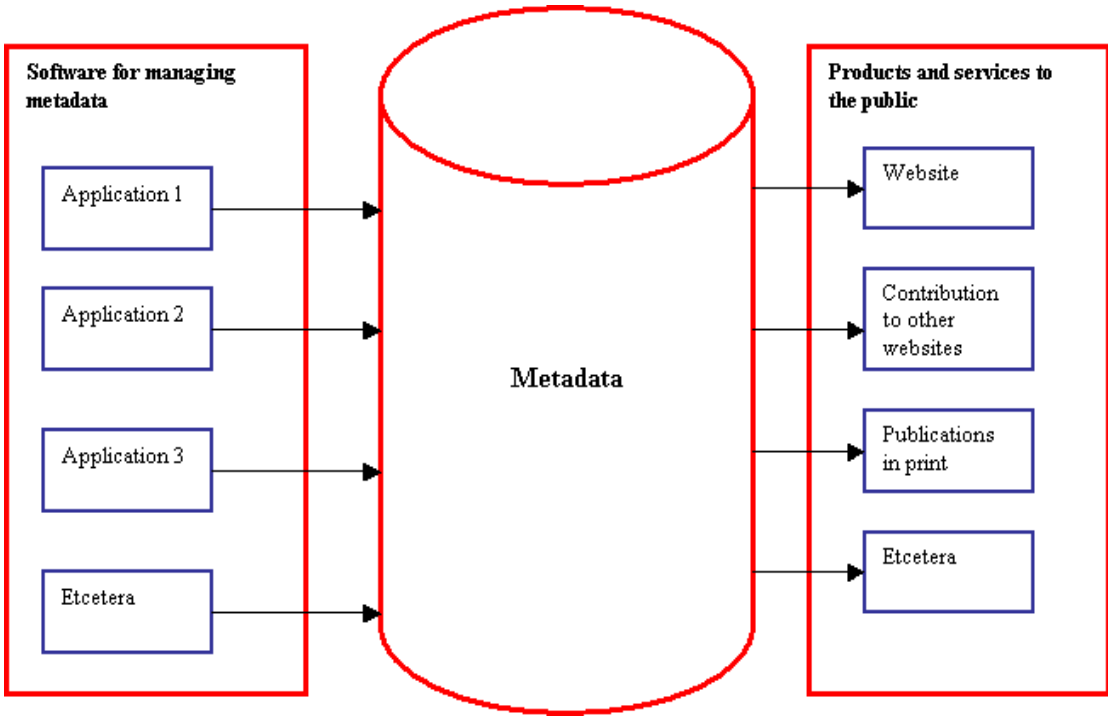

Diagram 4

Closing remarks: what is the use of IKM in practice?

The described models and concepts are useless when the implementation does not improve the performance of an organisation. And the proof of this improvement can only lie in the improvement of services to the users. Models like this can contribute to the definition of new projects and their impact on the organisation. They can also show the connection between various projects. It is a way of showing all departments in an organisation what the consequences are for all work processes if for example you decide to make a virtual exhibit on the Internet or to create an online search engine through metadata about your archives and collections. A model may help to reduce the risk of modifying work processes or metadata structures in such a way that they cannot be used anymore by other work processes. It also helps to prevent the introduction of new ICT-systems that will prevent the organisation from reaching goals like integration, standardization and presentation.

An IKM-model and an ICT-architecture provide a very effective framework for managing work processes and projects. Every work process and every project can be linked to each other. It is possible to make up a checklist for new projects to make sure that the results of the project fit into the preconditions of the organisation concerning IKM and ICT.

The most important assumption is that the management of the organisation should be aware of these instruments of improving services to the public. This awareness enables at the same time that all the work processes work together in such a way that these services can be created and maintained. It is up to management to create a policy for mid-term and long-term goals and objectives in such a way that projects can be started off in a way that is consistent with that vision.

To conclude this article, an example from the Municipal Archives of Amsterdam may be illustrative of an attempt to create new services to the public. The long-term aim is to create one single portal on the Internet to all descriptions of content and context of the archives and collections. This portal should be designed in such a way that the questions that are mostly asked by the users (questions about a person or organisation, about a subject, a location or a period) are answered in the simplest way possible. This is a huge task that will involve all work processes.

The starting point is very different from the situation that should be created. This is a fate that a lot of institutes now share. A lot of metadata is incomplete, not standardised, not kept in modern databases and not linked to each other. The Municipal Archives of Amsterdam not only hold archives but also vast collections of library material, audiovisual material, photos, drawings and maps. It is a huge task to integrate all the metadata related to this material, using appropriate standards. The work processes are not all structured in order to create integrated services to the public. In order to assemble metadata from

different sources a series of projects were defined. The first step was to create a complete set of accurate metadata for the highest aggregate level based on a survey of various access tools (finding aids). This survey is the first result of this [9]. Other search engines on the Web site are not yet linked to this survey. The next step is to present a set of metadata on lower aggregation levels for the archives. The material at present was not stored in a database. In order to realise this precondition a massive data-entry project was initiated. The result of this project is a database with 350.000 records. In order to present the metadata as an integrated service to the public it is necessary to enhance create and present indexes on persons, organisations, subjects, locations and time periods. This project has a twin brother in the back office, where existing ICT systems must be altered or replaced in order to create the necessary metadata in a standardised way.

In the next few years the intention is to integrate the metadata about context and content of all archives and collections into this model. The consequences are that almost all existing ICT-systems must be reconsidered, strategic decisions must be made about the choice of standards for metadata, a lot of conversion or data-entry of existing data must take place, quality controls and different ways of working should be implemented. This cannot succeed without defining an appropriate IKM-model and an ICT-architecture.

References

1. My special thanks go to Kent Haworth, York University Archivist and Head, Special Collections and Project Director and Secretary, ICA Committee on Descriptive Standards, for his comments on an earlier version of this article.
2. In most literature the abbreviation “IT” is used, I prefer to use the more modern abbreviation “ICT” - Information and Communication Technology.
3. Compare for example Marco, p. 5
4. Wurman, p. 27 and Milner, p. 3. I have replaced the word “wisdom” in this model with the more modest word “understanding”.
5. Compare for example Gilliland-Swetland, Anne J, *Defining metadata*, in Baca, p. 3;
6. Marco, p. 23-24. Marco cites Inmon, W.H.: Building the Data Warehouse, Wiley, 1996, p. 33;
7. Marco, p. 48
8. Applegate c.s., p. 139-140.
9. Survey URL: <http://www.gemeentearchief.amsterdam.nl/archieven_en_collecties/overzicht/introductie/index.nl.html>

Literature and suggestions for further reading

1. Applegate, Lynda, F. Warren MacFarlan en James L. MacKenney (1999). *Corporate Systems Information Management*. Irwin MacGraw-Hill, Boston.
2. Baca, Martha and others. (1998) *Introduction to metadata, pathways to digital information*. Getty Information Institute, New York.
3. Cook, Terry (2001). Archival Science and Postmodernism: New formulations for Old Concepts, in *Archival Science* (2001-1), ed. Horsman, P., E. Ketelaar and T. Thomassen, Kluwer Academic Publishers, Dordrecht, p. 3-24.
4. Getty Information Institute, New York, Art and Architecture Thesaurus URL: <http://www.getty.edu/research/tools/vocabulary/aat>
5. International Council of Archives (ICA), ISAAR(CPF) standard URL: <http://www.ica.org>
6. International Council of Archives (ICA), ISAD(G) standard URL: <http://www.ica.org>
7. Marco, David (2000). *Building and managing the meta data repository, a full life-cycleguide.* John Wiley & Sons, New York.
8. Menne-Haritz, A. (2001). Access: the reformulation of an archival paradigma, in *Archival Science* (2001-1), ed. Horsman, P., E. Ketelaar and T. Thomassen, Kluwer Academic Publishers, Dordrecht, p. 57-82.
9. Milner, Eileen M. (2000). Managing *Information and Knowledge in the Public Sect*or. Routledge, London. Records Continuum Research Group, Australia. URL: <http://rcrg.dstc.edu.au>
10. Ribeiro, Christina and Gabriel David (2001). A Metadata Model for Multimedia Databases.
11. Smit, F.P. (2000). Proposal for a Datamodel of Archival Descriptions, in: Atti del Summit DACE, Roma, 2000, p. 149-196.
12. Smit, F.P. (2001), Het nieuwe Overzicht van Archieven en Collecties, in: *Archievenblad* (2001-1), Koninklijke Vereniging van Archivarissen, Amsterdam, p. 26-29.

13. Society of American Archivists, Encoded Archival Description URL: <http://www.loc.gov/ead>
14. Svenonius, Elaine (2001). *The Intellectual Foundation of Information organisation*. The MIT Press, Cambridge Massachusetts.
15. Wurman, Richard Saul (2001). *Information Anxiety 2*. QUE, Indianapolis.

Frans Smit is Head of the Section of Archival Descriptions and Cataloguing at the Municipal Archives of Amsterdam. He is (and has been) also engaged in various national and international projects concerning providing access to metadata about archives and collections through search engines on the web.

- -

For citation purposes:
Smit, F. "The Historical Data Warehouse", *Cultivate Interactive*, issue 6, 11 February 2002.