European Commission

INSAR
Information Summary on Archives

EUROPEAN
ARCHIVES
NEWS

SUPPLEMENT VII

Proceedings of the DLM-Forum 2002

@ccess and preservation of electronic information: best practices and solutions

DLM-FORUM
Electronic Records

Barcelona, 6–8 May 2002

EUROPEAN
ARCHIVES NEWS
INSAR

- Secretariat-General of the
  European Commission

- Editorial Committee
  Supplement VII:
  Peter BERNINGER
  Frank BRADY
  Hans HOFMANN
  Jef SCHRAM

- For further information:
  European Commission
  Tel. (32-2) 29-50557
  (32-2) 29-56721
  Fax (32-2) 29-61095
  E-mail: archis@cec.eu.int
  Address: rue de la Loi 200
  Office: JECL 3/35
  B-1049 Brussels

Proceedings

# of the

DLM-Forum 2002

@ccess and preservation of electronic information: best practices and solutions

Barcelona, 6–8 May 2002

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Cataloguing data can be found at the end of this publication.

Luxembourg: Office for Official Publications of the European Communities, 2002

ISBN 92-894-4415-0

Printed in Italy

PRINTED ON WHITE CHLORINE-FREE PAPER

An expanding universe — metadata and accessibility of digital information

Johannes Hofman

Introduction

The increasing use of information technology (IT) has changed our way of dealing with information. On the one hand IT offers us new ways of creating, using, and making available information and on the other hand it also requires new approaches just because of that, and because of the different nature of digital information. Nonetheless, to date, a huge mass of digital information resources is available e.g. on the world wide web, and is growing every minute. Accessibility in this respect is a crucial feature of this digital information. How can that be achieved and above all maintained? To what extent are traditional tools and approaches still sufficient? That are questions that many organisations are facing to date.

In this respect the issue of metadata is all over the place. Everybody seems to have discovered this subject. Especially in the world of the world wide web people are getting increasingly concerned in information resource discovery and in better organising the overwhelming amount of information that is available. It shows the growing importance of the world wide web, but we have to be aware that this is not the only domain where information is managed and maintained. Business companies, memory organisations, such as libraries, archives and museums, government organisations and so on, they all create and manage huge information sources and they all have to deal with issues, such as how to maintain them and how to keep them accessible and understandable. The world wide web is in this respect in most cases 'just' a channel for distribution and dissemination, be it an essential one.

Two perspectives can be distinguished here and that is the perspective of the creator and custodian or preserver on the one hand and the perspective of the searcher or user on the other. Both perspectives have to be satisfied in order to be effective in the area of accessibility. Metadata plays a crucial role in it.

The approaches concerning metadata that are taken can be small or very broad in scope. They range from specific sets of metadata in a specific domain (e.g. geographical data) to a (de facto) standard for information resource discovery, such as the Dublin Core metadata set, to frameworks that help organisations to organise and manage their information sources and make them available and accessible as in the case of record-keeping metadata standards. These different perspectives and approaches show us the scope and also the underlying complexity of the issue. Another approach in this respect is the industrial or technical view. It takes the possibilities of IT as a starting point, and provides tools like automatic indexing, full-text retrieval, fuzzy logic, artificial intelligence etc. The issue is then how do they fit into the picture, how do they contribute, or what problems do they solve?

There is a plethora of initiatives, projects and ongoing research in different domains that are dealing with the issues described and that creates another problem. How to coordinate all these efforts, that do not only take place in one domain, but at the same time in many different domains and communities and from different viewpoints? The capability of Internet and e-mail in connecting people may be a resourceful instrument in overcoming that problem, but still action is needed to cope with it. Some cross-domain collaborative projects are already emerging and will help in exchanging information about new or other initiatives worldwide and in stimulating new research.

Nonetheless, it is necessary to identify how all these initiatives and their underlying questions and answers relate to each other. In the end it has to be the human being, in the quality of user, that should benefit from the results of all the work being done.

Hans Hofman

Hans Hofman (1948) has studied history, archival theory and informatics. Since 2000, he has been working as senior advisor for the government programme ´Digital longevity' on information management at the Ministry of the Interior (Department of Public-Sector Information Policy). This programme was initiated in 1996 by both the Ministry of the Interior and the Ministry of Education, Culture and Science, with the objective of creating appropriate conditions for electronic record-keeping within government organisations and archival institutions. In this position he is, among other things, involved in developing guidelines for digital record-keeping, metadata sets and in formulating strategies for digital preservation.
On the international scene he is (since 1993) a member of the Committee on Electronic Records of the International Council on Archives.
Within the European Union he is a member of the DLM-Monitoring Group and as such involved in activities to stimulate cooperation in the field of electronic records. He is also co-director of the recently started European project Erpanet (electronic resource preservation and access network) on digital preservation.
Furthermore, he is investigator and representative of the National Archives of the Netherlands, participating in an international research project initiated by the University of British Columbia (UBC), called the InterPARES research project (1998–2001), which has the objective to investigate the long-term preservation of the integrity of electronic records.
He is since 2000 representative for the Netherlands in the TC46/SC11, working on ISO Records Management Standard 15489.

In this paper I will try to identify the main issues relating to metadata, the different perspectives taken in different domains and related projects, to some extent discuss how they relate to each other, and what possible approach can be taken. The challenge is to achieve better coordination and to identify ways forward.

What are the issues? Perspectives and purposes

In order to be able to understand the complex area of metadata and the role they play, it is necessary to identify what the (main) issues are and what is being pursued in all these different projects. In general one could say that depending on their point of view these efforts aim at discovering, disclosing and retrieving information, at enabling understanding and interpretation, or thirdly at enabling management and preservation.

The essence of metadata is that they provide us with the necessary information to understand and use information. It starts for example already with the need for communicating. An e-mail message needs information to whom it has to be sent or from whom it is coming, and what the subject is. We may also want to send a copy to somebody else. As soon as we receive a message we want to keep information about when it was received (date and time), and whether it was a reply or just a first message. Other information that may need to be kept is, where we have stored it and under what number and/or name, what happened since then (did it stay unchanged), how it is related to other messages or other documents, and so on. It is all metadata.

Some people say that a reason for collecting metadata about digital objects is, that it is easier to handle a small set of metadata than the objects themselves. That is one argument, but there are other more important principles.

To be able to use information resources it is necessary to satisfy at least three basic requirements. These requirements are: to be able to find the information, subsequently to be able to interpret it and finally to know whether the information can be trusted or not ([1]). Depending on the domain in which it is created and used these requirements may differ in strictness. In an organisation that is vulnerable to lawsuits requirements for authenticity and reliability of information or records will be severe, but in case information is retrieved for more informal reasons the requirements of authenticity and integrity will be less, but those for retrievability, interpretability and meaningfulness will still be important.

It is clear that in relation to the world wide web much attention is being paid to information discovery and retrieval and with reason. After all one of the main goals is to find and use information. Retrievability however only makes sense if all the basic requirements are satisfied. In the following section the basic requirements are further characterised.

Retrievability

The main thing is to be able to find information, but above all information that is relevant to you, or that answers your question. That is the ultimate goal of the work being done in all kinds of organisations that make information resources available.

Searching information on the world wide web is enabled through search engines, which allow people to enter keywords representing subjects they are interested in, hoping they will find information resources that will satisfy their needs. The well-known issue here is the huge amount of hits that are in most cases returned or provided to the user. Although there are mechanisms that value the information resources to what extent they comply with the keywords and arrange the information sources accordingly, there are also many mechanisms that arrange the sources according to other criteria, such as the amount of links or the occurrences of keywords on web pages. The amount of tricks and deceptive methods used here is huge however and makes one suspicious about the retrieval results. Some people applaud or rely on the possibility to search on the content of documents (full-text retrieval) as an easy and useful way of finding the right information. They easily forget however that the use of words in documents is not controlled by any mechanism, except for the human mind, and that has proven to be very unreliable and inconsistent in this respect. So using this method offers hardly any consolation, but probably more frustration. Perhaps in the best case it may be used as an additional tool for retrieving information resources.

[1] In a document, written in 1999 by the *Description and classification of government records Working Group* of the *Information Management Forum,* in Canada these requirements are also mentioned as the aims for archival description. The document is titled: 'Approach to the description and classification of government records'. (www.imforumgi.gc.ca/new_docs/draft_e.html).

As such this is not a really effective way of searching. Which requirements in relation to retrievability should be satisfied? Essential is that objects have to be identifiable and to be located. A very important issue in this respect is the unique identifier that has to be persistently linked to the object. The uniqueness is dependent on the domain in which objects will be used and can or should be extended to all domains, certainly if we talk about the world wide web. Mostly an identifier will be unique only within the domain in which the document or information object has been created, not for other domains. The well known problem with the current identifier, i.e. URLs is that they are not reliable for finding information, since they are location-oriented (e.g. web domain name). When that location changes, and that happens often, the web address changes. Efforts are being made to solve this problem. Examples are URN and DOI initiatives. Whether they will succeed, is still a question.

Apart from being locatable it is necessary to structure or organise and describe the information objects. Different methods are available for doing so. Examples can be found in libraries (mostly subject-based, e.g. UDC) and archives (mostly function- or activity-based, e.g. business classification schemes). The structure establishes intermediate levels which allow better navigation and make retrievability easier.

Understandability

As soon as the information object has been found and presented, one has to be able to understand and interpret it. Otherwise we cannot value the content in relation to the question we have. That means that information about the origin has to be available, e.g. why and how has it once been created and used, does it have a relationship with other documents or information objects and so on. It might be that not all of this information is necessary for each question, but it has to be available. This meta-information (metadata) has to be described and linked to an object. Issues that may play a role here are, is it written in a language we understand, do we understand enough about the background of the object — which issue may become very important if the object has been created a long time ago — or is the information resource coming out of another domain of knowledge or activity, that we are not familiar with. It all requires metadata that has been captured at its creation or may be added afterwards. These metadata can be embedded in the object itself or can be external or both.

Trust

Finally if we have found an object and are able to interpret it, we still need to have a feeling about the trustworthiness of the presented information. In a digital world this has become even more necessary, because everything in cyberspace is information originating from all kinds of sources, and worse, because digital information has no fixed form and as such is very volatile and vulnerable to alteration or mutilation. In the case of the world wide web this is aggravated by the fact that it is not a controlled environment. So the information has to be taken care of continuously behind 'the screens'. If not, it will be difficult to prove that it has not been tampered with. In some cases that may not be an issue, but in many cases it will, especially in the case of lawsuits or research for instance.

The history of how the object has been managed (can be e.g. an audit trail) is essential for proving that the information presented, is not corrupted and therefore trustworthy. Proper management requires all kinds of procedures, methods and measures that ensure a safe environment for digital resources. In the InterPARES project efforts have been made to identify the requirements that contribute to or provide trust and how these should be implemented in preservation systems in order to achieve the desired outcome: trust in the results presented based on the search and reproduction methods. Tools as digital signatures can be used as additional mechanisms that help to ensure this trustworthiness. Several other initiatives are under way to identify attributes and functionality for ensuring reliability of information, such as the ISO Records Management standard and the RLG report on 'Attributes of a trusted digital repository' ([2]).

The basic requirements as discussed do not fundamentally differ between different domains. The emphasis may be different or the strictness of the requirements, but in fact all three sets of basic requirements have to be accomplished always. Where in records management or archival com-

[2] RLG/OCLC: 'Attributes of a trusted digital repository. Meeting the needs of research resources', draft for public comment (August 2001). See www.rlg.org/longterm. ISO Records Management Standard 15489 represents in itself requirements for such a trustworthy environment. The InterPARES project as mentioned has gone into its second project (2002–07) to push issues around authenticity and integrity especially of records further and to provide guidance (www.interpares.org).

munities these goals or compliance to these requirements are being pursued in carrying out records and archival management, in other domains such as libraries this management area is mostly called (digital) 'preservation'. The difference is that the latter focuses on the continuing usability and availability of digital information objects, being mostly electronic publications and more and more web resources. Record-keeping or records management is the discipline that deals with records from their creation on for as long as they are needed, which can be for 'eternity'. As such it can be considered to have a broader scope. Appraisal and describing the interrelationships between records are core activities.

At the moment these approaches seem to be more or less competing with each other, because communication between the two communities does not really exist, only on ad hoc basis. There is a predominance of the library community. Whatever the reason for this may be, in order to improve information exchange or cross domain searching it is necessary to find ways to identify the commonalities, to achieve more synergy by using skills from different domains, and see how these can be used as a basis for further research and development. Besides, there are other communities, such as research institutions with scientific data, cultural heritage or industry (e.g. pharmaceutical industry) trying within their own domain to find solutions as well. More openness and information exchange in this area is needed, in order to learn from each other. That is also necessary for another reason, because apart from the above mentioned basic but essential requirements there is another relevant issue, called interoperability. In the openness of a networked environment such as the world wide web it is necessary to coordinate the efforts to improve communication and exchange of information between domains. However there is the issue of different semantics. A publisher in a library environment uses a different terminology than people in public administration or e-government for instance. It is the unavoidable problem of similar terms in different contexts for different concepts or of different terms for similar concepts. These different domains have each their own perspective and domain bound terminology and that will lead to different metadata sets. How to reconcile the different interests or perspectives? What solutions or approaches are available?

In the different domains where this issue is addressed, there is a growing awareness of this interdependence. So initiatives are developed to learn about what is happening elsewhere and to develop instruments that enable interoperability. Examples are for instance the Harmony project that aims at achieving semantic interoperability between different sets of metadata in e-publishing so users can search (electronic) publications on the web uniformly, and the European Schemas project that tries to establish an information service that provides information about the different metadata schemas that are developed and how they relate to each other. It also tries to map them ([3]).

Both projects build heavily on Dublin Core developments. It is remarkable, however, that developments in the area of records management or record-keeping until now hardly have been taken into account. That might have two reasons: there is nothing to report on in this area and/or what is there does not fit the needs of the community or at least that is what people think. A third reason may be the fact that this area is not well known to the outside world. I guess it is a combination of the latter two.

Longevity or preservation

There is another dimension, if we extend interoperability in time. It will mean that we have to take care of the ongoing (technical) readability and the meaningfulness of the information objects and that concerns the area of preservation or management of information resources over time. Both aspects of preservation, intellectual and technical, have to be addressed and need permanent maintenance in order to keep the information resources involved accessible and understandable. The same goes for the meta-information about them.

Theory, methods and practices

The next step will be how to implement the abovementioned requirements. Summarising the above paragraph the following aspects and activities can be distinguished:

(1) description of the information resources (either publications or records or data sets);

[3] See for Harmony project www.ilrt.bris.ac.uk/discovery/harmony and for Schemas www.schemas-forum.org and www.cultivate-int.org/issue3/schemas.

(2) persistent identification through time and across domains;

(3) ongoing contextualisation to provide meaning to information sources through time;

(4) interoperability between different sets of metadata used in different domains;

(5) standardisation at different levels (as regards to e.g. structure, semantics, value and/or content).

Although these aspects and areas of research are not mutually exclusive and the overview is not exhaustive, it shows the complexity of it.

Instead of discussing all these issues in detail I like to take a slightly different approach and try to identify a core set of activities to make and keep information accessible. Basically the following instruments or methods can be distinguished:

- Appraisal and selection of what will be preserved. This is always necessary, in order to preserve no information that is not needed and has no value. Inherent to this is the necessity of disposing of information as soon as it is no longer needed (clearing up).
- Structuring or clustering of information resources according to certain criteria. In the case of records this structuring activity is done mostly based on business needs by using for instance classification schemes and has two reasons:
  — to cluster the documents or any other information entities that are interrelated (in business processes these documents are evidence how a certain case has been handled for instance). That can be done almost automatically, since the documents reflect a business process and the way a case is processed. The objective is to articulate or express the documentary context and to maintain the coherence between documents ([4]);
  — to cluster information around subject/business activities in an organisation that created the information. This is mostly done at a higher level. In libraries structuring takes place mostly based on subject classification.
- the third main instrument is 'describing' the information objects by 'labelling' them, e.g. through a classification scheme, and by making them identifiable. For records this description can be largely derived from the activity that creates them. As indicated one of the main areas for adding metadata is also to provide contextual information on information resources as in the case of records. That makes it possible to understand and interpret them. These metadata provide information about the origin or provenance, nature, state, content, structure and access of an object so we can establish or assess the value, reliability, authenticity etc. Essential is to keep the resources meaningful through time.

In both the record-keeping and archiving community and the library community these methods are well known and used since long. The application of them will be different because of the different nature of the material in custody.

Different approaches are under way in order to deal with these issues. It is hard to distinguish the many different initiatives in different domains and with different perspectives and to keep up with new developments. As already indicated they take place mainly in the area of information-resource discovery. Some of these approaches are object oriented, some are process-oriented, some are function oriented or even a combination of these.

Examples of object-oriented approaches are for instance information resource discovery initiatives such as Dublin Core, as well as the European MiReG project (managing information resources for e-government) ([5]). Function-oriented approaches are the Open Archives Initiative (OAI), trying to achieve interoperability in the publishing area, or Cedars, the research libraries group (RLG), and the open archival information system (OAIS) in the preservation area ([6]). Others are the ISO Records Management Standard 15489 and its consequences for metadata, and the archiving metadata forum (AMF) in the area of record-keeping, though they may also be called (business) process oriented ([7]).

One of the instruments that is being introduced to serve as a describing mechanism, is the Dublin Core metadata set which offers some control on retrieval of information resources. It allows to add information to a document, publication or web page that is not within the document itself for

([4]) This is called the 'archival bond'. See Luciana Duranti, www.slais.ubc.ca/users/duranti.

([5]) MiReG: http://ag.idaprog.org/Indis35prod/doc/312 and http://dublincore.org/groups/government/mireg-metadata-20010828.shtml.

([6]) See for OAIS: http://ssdoo.gsfc.nasa.gov/nost/isoas; Cedars: www.leeds.ac.uk/cedars; RLG: www.rlg.org/longterm.

([7]) ISO RMS 15489 is produced by ISO TC46/SC11. Archiving metadata forum: see www.archiefschool.nl/amf.

instance. In the world of the Internet it seems as if the Dublin Core metadata set is the only tool available. With this 'hammer' some people try to make a 'nail' of everything. Notwithstanding its merits and the fact that it has been adopted worldwide as a *de facto* standard for information resource discovery, it has however its limitations as a publication and resource discovery-based instrument. That is shown for instance in attempts to extend this standard with new elements to satisfy other requirements, such as in the case of the Australian government locator system (AGLS), where four elements are added to the original 15 DC elements. Unfortunately there exist also ideas to extend this set with record-keeping elements. This approach is completely denying the complex and different nature of record-keeping metadata.

A rather recent and interesting development is the emergence of the semantic web and with it ontology. Ontologies describe and structure terms of a certain domain of knowledge into a controlled vocabulary, and as such structure the metadata into hierarchical structures/classes, subclasses etc. It provides a tool for identifying the relationships between terms in that domain. This approach of adding semantics to information resources will enable intelligent agents to search much more efficiently and effectively.

Apart from these more sophisticated approaches industry provides as indicated tools such as automatic indexing. They promise to make things easier, but the question is do they really? Such techniques are dependent on the use of words and terms in documents and as can be expected that is not consistent. The result of automatic indexing, how intelligently it may be done, can never be as good as deliberate and structured metadata creation, capture and management.

An important issue in this respect is to ensure that adequate metadata is generated at the moment of creation of the document or source, that is captured and maintained in a useful way and persistently linked to it. A proactive approach is much more efficient than a retrospective one, in which case information resources are labelled or 'manipulated' when they already exist. The proactive way might be 'easier' or perhaps more natural to achieve in a business activity environment than a more open environment as for instance that of publishing (books or web pages), but it is essential from a cost-effective point of view.

If the requirements for metadata are clearly identified, it will be possible to develop software tools that in the case of records will enable automatic capture of these metadata from the business system with which the records were created, itself or from closely related systems, such as workflow systems. By integrating metadata capture in software applications the cumbersome task of gathering metadata that will give meaning to information sources of whatever kind will be more easily accomplished.

In this respect it is interesting to see what is happening in practice, for example in e-government initiatives. Several governments are now trying to establish standards and frameworks for metadata, in order to make interoperability between government organisations possible. These standards so far focus mainly on information resource discovery. However there are also some interesting examples to establish and include recordkeeping metadata sets for government agencies, such as in Canada, Australia and the United Kingdom. Each of them tries to identify a minimum set that should guide government organisations in managing and maintaining their records.

The Canadian 'record-keeping metadata requirements' are produced in January 2001 and consist of a minimum set of 26 metadata elements that allows organisations to describe and share information and meanwhile facilitates interoperability ([8]). Eleven of these elements overlap with the Dublin Core metadata set. The Australian Record-keeping Metadata Standard for Commonwealth agencies (RKMS) consists of 20 elements of which eight are mandatory. These sets of elements however are very high level and need further refinement and explanation with sub-elements and qualifiers to be implemented and used.

Government organisations are free to add these specific elements or sub-elements. In general the Australian minimum set consists of three parts, describing the organisation, the document or record and the management history of the record respectively. The set however is focusing on records, and not describing the full context of it ([9]). With this set though agencies at least know what metadata they should capture in their record-keeping systems.

[8] See for the Australian RKMS: www.naa.gov.au/recordkeeping/control/rkms/summary.html and for the Canadian set www.im-forum.ca.

[9] The set is currently being review in order to improve and adapt it based on the experiences so far.

Although the RKMS intends to describe the records and is not focusing on retrieval of records it is the idea that it should be in line with information resource discovery metadata sets as much as possible. In the case of the Australian RKMS a strong relationship and overlap exists with the abovementioned Australian Government Locator Service (AGLS). This set is accompanied by the Australian Government Interactive Functions Thesaurus (AGIFT), that is an addition to the function element of the AGLS standard and provides a controlled vocabulary for describing government functions, which creates a strong link between information resource discovery and record-keeping ([10]). This approach reflects awareness that it is necessary to link both worlds, a view which sees resource discovery metadata as a subset of recordkeeping metadata.

The third example is the 'E-government framework for metadata' in the United Kingdom as published in 2001. It recognises the need for standards to ensure consistency in effective information management and intends to provide a framework for government organisations for dealing with resource discovery and records management. The main objectives are to enable effective search through metadata instead of the resources themselves and to make people confident that the retrieved source being presented is the best one. The framework introduces the Dublin Core as the accepted standard, though admitting that this will not be sufficient, and envisages also the development of a rather ambitious pan-government thesaurus.

These examples show the increasing awareness that a broad metadata framework is necessary, which includes both information resource discovery and management metadata. It will not only make better communication between government organisations possible, but also if properly implemented compliance with requirements on trust and understandability.

The fact that metadata tags, such as provided by the Dublin Core set, are meant for information resource discovery, makes it possible to identify the possible overlap with record-keeping metadata, which are mainly focusing on the enduring interpretability, authenticity and integrity of a specific set of information resources, namely records.

Towards a common framework for metadata

The basic notion that can be derived from the previous paragraphs is that two different, seemingly contradictory viewpoints can be distinguished:

(1) the needs of a user or (re)searcher, including reuse of information resources for other purposes than they were created; and

(2) the need of managing and maintaining information sources in order to keep them trustworthy and understandable.

It also seems as if different communities are taking care of each of these perspectives, i.e. the library community on the one hand, and the records management and archives community on the other. This is a general picture of course, which is not completely justified by practice. Apart from that within these and other communities many different metadata sets or standards exist, which make it even more complicated. One of the things necessary is to map these sets and see how they can be connected. It adds another meta-level, but it is an illusion to think that there will be one common metadata set shared by all communities. The consequence is to identify at what level the existing sets can communicate with each other and develop a conceptual framework for it. An example of such an initiative is the ABC-model that is being developed in the Harmony project ([11]). It identifies a set of entities that is common in many metadata sets in different domains and intends to provide a general logical model that describes them and their interrelationships. These entities regard people, organisations, places, events etc. The model focuses on events and the basic idea behind it is that information resources can evolve or transform over time by events, e.g. a translation into another language of an information resource. That event then influences the description of the resource, because one or more properties of the resource have changed. Events are connected to agents, dates, places etc.

This approach is interesting because there is a parallel to the creation of records. Records are the results of activities carried out by agencies in doing business. A metadata model based on that notion can be found in the SPIRT-model, as developed by Monash University. That high-level

([10]) See also Cunningham, A., 'Six degrees of separation: Australian metadata initiatives and their relationships with international standards', *Archival Science* (Vol.1, No3, 2001) pp. 271–283.

([11]) See Lagoze, C., Hunter, J., Brickley, D.,'An event-aware model for metadata interoperability', 2000.

model identifies three basic entities, agents, business and records, that can have all kinds of relationships ([12]).

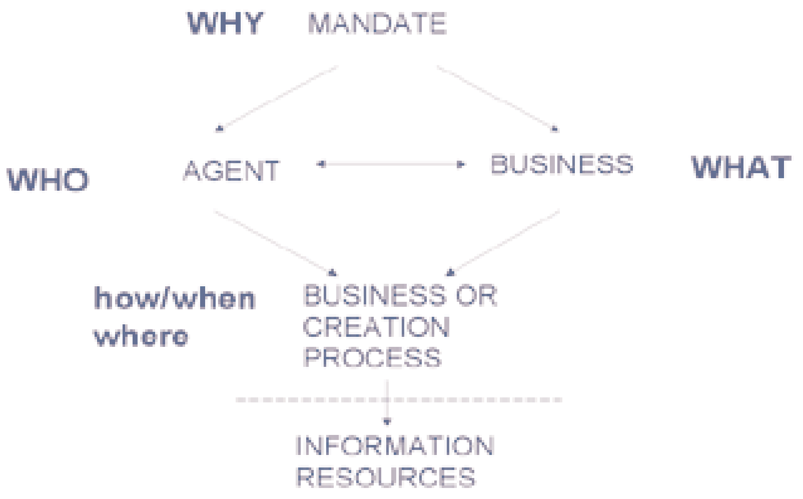
The basic scheme that follows out of the previous models and remarks is that people carry out activities or do business which results in information resources (records or publications). This perspective is taken especially by the SPIRT model. The perspective taken by ABC-model is that information resources can or will be transformed by events, carried out by agents. Despite these different viewpoints there is a strong overlap in entities. And there is a third perspective, the viewpoint of the researcher or user. The question asked by the user is mostly based on who did what when, where and/or why? As such this question or part of it can easily be related to information resource creating activities.

The following diagram shows at a high and simplified level the relationships between the above mentioned entities or elements.

What we need to know in order to be able to fulfil the requirements of retrievability on the one hand and of interpretability on the other hand can be identified as rather simple: who, what, when, where and in some cases why. The 'who did what why' provides us with the information about provenance and identity. At the level of the information sources itself (publications, web pages, or records) metadata on their management (activities such as appraisal, maintenance, description, access etc.), preservation and use have to be captured. The elements of when and where or time and space are applicable to the whole, because agents or organisations and activities including their interrelationships will change over time.

A similar approach is for instance used in the electronic resource citation (ERC) ([13]). This citation idea is based on Dublin Core elements and intends to provide a metadata kernel with a very simple format (who, what, when, and where), that should support the permanence of network discoverable objects. The approach though is rather static and does not take into account the dynamics of metadata description.

Figure 1: Basic entity model



Nonetheless there seems to be a common basic scheme that can be used for different purposes. If we are able to build a model around these elements, we will have a solid core set and the most important needs can be satisfied. This can in principle be applied to all kinds of information sources.

Concluding remarks

The globalisation of the information world has a strong impact on the way how we manage it. The emergence of the world wide web requires new approaches and methods in order to enable easy access and accessibility. Metadata play a key role in this. At the moment there is a predomi-

[12] See for the SPIRT model, for instance, www.sims.monash.edu.au/rcrg and McKemmish, S., Acland, G., Ward, N., and Reed, B., 'Describing records in context in the continuum: the Australian record-keeping metadata scheme', *Archivaria* 48 (Autumn 1999).

[13] Kunze, John A., *A metadata kernel for electronic permanence*, JoDI (Journal of Digital Information), A special issue on metadata: Selected papers from the Dublin Core 2001 conference, (Vol. 2, Issue 2, January 2002). See: http://jodi.ecs.soton.ac.uk/articles/v02/i02/.

nance of approaches from the library community. The tendency or even the need, based on the globalising effects and the openness of the world wide web, towards better coordination and collaboration between different information providers and communities requires other attitudes in these communities. This is especially in the archival community achieved rather slowly.

In order to be present, active and effective in this new virtual world it is necessary to be aware of the characteristics of it. In the area of metadata the world of information resource discovery has to be better linked to the world of management and preservation of information resources. This means that sets of metadata have to be mapped with each other, but also that there has to be a better understanding of the different perspectives that exist. Only then will it be possible to achieve the required interoperability. As indicated there are some promising initiatives in this respect.

One question in making information resources accessible in a digital environment is, do we need metadata or should we make use of the possibilities of IT or software to search the content of these sources? Isn't one of the big benefits of IT that computers made search on content of documents possible, while before it was not? Obviously, looking at what is happening at the moment as regards all existing metadata initiatives that seems not to be sufficient at all. The arguments used are the need for reliability, interpretability and interoperability of information resources of all kinds. In this respect the idea of automatic indexing is not relevant, because it does not address these requirements. Moreover, this tool is insufficient in order to deal with the different sets of words, the different semantics, the inconsistency in the use of words etc. in information resources and as such it does not contribute to or solve the issues mentioned. At the most it may provide an additional help. The same goes for other tools as full text retrieval or fuzzy logic tools.

Guidance is needed to find one's way through the dense forest of existing projects and initiatives around metadata in all its forms, and to understand how they relate. In this paper I have tried to identify and describe a possible common concept that could be a solid basis to build on. This concept is in line with the needs of both the creation and the use of information resources.

It is also clear that there are more needs to be served than information discovery and more approaches possible or available that may play a role on the scene of retrieving and managing information. Especially one discipline may bring relevant and useful experience and approaches to other communities and that is the archival community, which for centuries has been and still is very familiar with managing, preserving and making information sources accessible. Although common knowledge among records managers and archivists, it becomes more and more obvious that outside this specialist and small community not many people are aware of that.

So on the one hand there is a community that focuses on information resource discovery and is seeking for a common instrument for making information resources retrievable and searchable by establishing a limited set of tags. On the other hand there is the archival community (including records management) that captures, organises and manages a specific category of information, being created in the course of doing business, called records.

They are complementary, and bring different but relevant skills to the floor. Both perspectives are necessary to help people to retrieve information easily, to assess what the information is about, and whether they can trust it and interpret it. They meet in the domain of the world wide web, but they still have to connect properly, so synergy can be achieved from what they each are trying to do. Coordination can be improved. The same goes for collaboration with software suppliers that can provide useful tools based on the identified requirements.

This paper I hope also makes clear that adding metadata is not a useless burden, though it has to be cost-effective and user-friendly. One could see metadata as the 'value added tax' (VAT) of information. It may be a costly thing and it may be experienced as a burden, but more importantly, it has or should certainly provide added value. So it would be better to speak of metadata as 'value adding tags'. It is up to the specialists, both information and IT, to make it easier e.g. by the use of IT.

Un universo en expansión: metadatos y accesibilidad de la información digital

Johannes Hofman

Al desarrollar un mundo empresarial y de administración electrónica, creamos, comunicamos y utilizamos cada vez más información digital. La ofimática, el correo electrónico e Internet son herramientas potentes a este respecto, pero recogiendo las palabras de Tim Berners-Lee: «[...] las herramientas potentes pueden muchas veces utilizarse con fines constructivos y destructivos [...]». Estas herramientas pueden pues utilizarse para bien y para mal, y se trata de saber cómo obtener el máximo beneficio de ellas.

Otro aspecto interesante que debe tenerse en cuenta es que las actividades vinculadas a la información convergen en uno único mundo virtual, la red, a la que se accede desde la oficina (es decir, desde la pantalla del ordenador). En otros términos, las tecnologías de la información ofrecen medios radicalmente diferentes de registrar, manipular y utilizar la información, pero por otro lado exigen nuevos enfoques para gestionar y tratar esta información. Ahora bien, no siempre administramos adecuadamente la información digital. La pérdida de información es uno de los riesgos de la evolución actual, no sólo como tal, sino también porque no somos conscientes de la existencia de este riesgo. En consecuencia, no sabemos detectarlo cuando es necesario. Debemos por tanto gestionar la información digital y su accesibilidad.

Los medios tradicionales de investigación y acceso a la información no son suficientes en un mundo digital, habida cuenta de la naturaleza diferente de esta información. Contrariamente a los documentos clásicos de papel, los documentos o la información digital ya no son entidades materiales fijas, sino que son intangibles y volátiles. Necesitamos pues nuevos mecanismos adaptados y complejos para hacerlos visibles, recuperables, accesibles y comprensibles.

Varios proyectos y enfoques se esfuerzan actualmente en cumplir estas condiciones. Algunas profesiones, en particular los bibliotecarios y los archiveros, se ocupan básicamente de la utilización de metadatos de todas clases, mientras que los fabricantes de *software* proponen herramientas, tales como la indexación automática y la búsqueda en el texto íntegro. Otros instrumentos son los diccionarios y las normas. La cuestión radica en saber qué enfoque es el más útil, en qué circunstancias y con qué objetivo.

Para poder utilizar las herramientas, es necesario entender la finalidad de su utilización, por ejemplo: el suministro de servicios electrónicos, el acceso a la información, a la comunicación, a la conservación, a la gestión, etc. Todas estas actividades tendrán requisitos diferentes, eventualmente con algunos solapamientos. Cabe distinguir al menos dos perspectivas: la perspectiva de la creación y del mantenimiento de la información y la perspectiva de su utilización. Si el creador y el usuario no son una misma persona, su visión será diferente y por tanto también lo serán sus necesidades.

La red implica también la utilización de una lengua común de investigación y comunicación. La interoperatividad de la información en los distintos ámbitos y competencias es cada vez más importante. Aparte de la interoperatividad, la posibilidad de recuperar datos en sí no es suficiente. La información recuperada y encontrada debe presentarse y probarse en cuanto a su fiabilidad, y como último aspecto, pero no menos importante, debe comprenderse para poder interpretarse. Todos estos aspectos influirán en la gestión de la información digital y en su accesibilidad.

Existen numerosas iniciativas que tratan de encontrar soluciones para buscar información en la red de manera adecuada. Estas iniciativas están dirigidas en su mayoría por las necesidades de una comunidad, como es el caso de la iniciativa IAA (Iniciativa de Archivos Abiertos) o del proyecto DCMI (Dublin Core Metadata Initiative). El desarrollo de la «red semántica» constituye otro ejemplo reciente.

Por lo que respecta a la evolución de la comunidad archivística, la codificación de la descripción (EAD) y del contexto (EAC) de los archivos son otros medios de presentar información archivada a comunidades de usuarios. Estas soluciones se basan en series de metadatos de archivo existentes.

La exposición mencionará los aspectos que rodean a la recuperación y la accesibilidad de la información digital, en particular la creación de metadatos, la interoperatividad y los principales aspectos de la evaluación de los enfoques, las iniciativas y las herramientas actualmente disponibles.

Ein Universum im Wachstum: Metadaten und Zugänglichkeit digitaler Informationen

Johannes Hofman

Im Zuge der Entwicklung einer Welt mit elektronischem Geschäftsverkehr (E-Business) und elektronischer öffentlicher Verwaltung (e-Government) schaffen, verbreiten und nutzen wir eine immer größer werdende Menge digitaler Informationen. Büroautomatisierung, E-Mail und Internet sind dabei starke Werkzeuge, doch wie Tim Berners-Lee sagte, lassen sich starke Werkzeuge oft sowohl für konstruktive als auch für destruktive Zwecke benutzen. So können sie entweder zum Guten oder zum Schlechten dienen, und die Frage ist, wie wir das Beste aus ihnen machen können.

Ein anderes relevantes Problem besteht darin, dass alle informationsbezogenen Aktivitäten in einer virtuellen Welt zusammenlaufen, und zwar im Internet, zu dem man vom Schreibtisch (d. h. vom Computerbildschirm) aus Zugang hat. Anders gesagt, ermöglicht die Informationstechnologie vollkommen andere Wege der Aufzeichnung, Manipulierung und Nutzung von Informationen, erfordert andererseits aber auch neue Konzepte zur Verwaltung und Handhabung von Informationen. Es sieht doch so aus, dass wir noch immer nicht in der Lage sind, digitale Informationen sachgerecht zu verwalten. Eine der Gefahren dieser aktuellen Entwicklungen ist der Verlust von Informationen, und zwar nicht nur der Informationen an sich, sondern auch, weil wir nicht wissen, dass sie vorhanden sind. Dadurch können wir sie nicht finden, wenn wir sie brauchen. Es geht also um das Management digitaler Informationen und ihrer Zugänglichkeit.

In einer digitalen Welt wird die herkömmliche Vorgehensweise beim Retrieval und Zugriff auf digitale Informationen wegen des andersartigen Charakters dieser Informationen nicht ausreichen. Im Gegensatz zu traditionellen Papierdokumenten sind digitale Dokumente bzw. Informationen keine festen physischen Einheiten mehr und daher flüchtig und nicht greifbar. Daher benötigen wir neue, zweckentsprechende und moderne Mechanismen, um sie sichtbar, recherchierbar, zugänglich und verständlich zu machen.

Momentan versucht man mit vielen Projekten und Konzepten, diesen Erfordernissen gerecht zu werden. Fachbereiche wie das Bibliothekswesen und das Dokumenten- und Archivwesen beispielsweise befassen sich in erster Linie mit Metadaten aller Art, die Softwareindustrie hingegen bietet Tools wie automatisches Indexieren und Volltextsuche an. Andere Instrumente sind Thesauri und Normen. Dabei stellt sich nun die Frage, unter welchen Umständen und zu welchem Zweck welches Konzept – wenn überhaupt – am hilfreichsten ist.

Um Tools nutzen zu können, muss klar sein, wofür sie verwendet werden, z. B. elektronische Dienstleistungserbringung, Zugriff auf Informationen, Kommunikation, Archivierung, Verwaltung usw. Jede dieser Aktivitäten ist mit anderen, sich vielleicht zum Teil überschneidenden Anforderungen verbunden. Es lassen sich mindestens zwei Sichtweisen unterscheiden: die Sichtweise der Erstellung und Pflege und die Sichtweise der Nutzung von Informationen. Handelt es sich beim Urheber und beim Benutzer nicht um dieselbe Person, sind ihre Sichtweisen und damit auch ihre Anforderungen unterschiedlich.

Das Internet erfordert auch eine gemeinsame Sprache für die Recherche und die Kommunikation. Daher gewinnt die Interoperabilität von Informationen in den verschiedensten Zuständigkeitsbereichen und Domänen zunehmend an Bedeutung. Abgesehen davon wird Recherchierbarkeit allein nicht ausreichen. Die gesuchten und gefundenen Informationen müssen auch dargestellt und auf ihre Vertrauenswürdigkeit hin geprüft werden, und nicht zuletzt muss man sie verstehen, um sie interpretieren zu können. All diese verschiedenen Gesichtspunkte machen es erforderlich, dass im Hinblick auf die Verwaltung digitaler Informationen und die Zugänglichkeit etwas unternommen wird.

Es gibt zahlreiche Initiativen, die sich um Lösungen für das Auffinden sinnvoller Informationen im Internet bemühen. Dahinter stehen zumeist breite Benutzergruppen, wie beispielsweise im Falle der Open Archives Initiative oder des Dublin Core-Standards. Ein anderes interessantes aktuelles Beispiel ist die Entwicklung des Semantic Web.

Im Dokumenten- und Archivwesen sind Entwicklungen wie die Encoded Archival Description (EAD) und der Encoded Archival Context (EAC) weitere Mittel zur Darstellung aufgezeichneter Informationen für Nutzergruppen. Sie basieren auf vorhandenen Metadaten der Aktenführung.

In diesem Beitrag werden Retrieval- und Zugriffsmöglichkeiten bei digitalen Informationen erörtert, so etwa Erstellung von Metadaten, Interoperabilität und Schlüsselfaktoren für die Bewertung vorhandener Ansätze oder Initiativen und Tools auf diesem Gebiet.

Un univers en extension: métadonnées et accessibilité de l'information numérique

Johannes Hofman

En développant un monde d'e-business et d'e-administration, nous créons, nous communiquons et nous utilisons un nombre sans cesse croissant d'informations numériques. La bureautique, le courrier électronique et l'internet sont des outils puissants à cet égard, mais, pour reprendre les mots de Tim Berners-Lee, «des outils puissants peuvent souvent être utilisés à des fins constructives et destructives»… Ces outils peuvent donc être utilisés pour le meilleur et pour le pire, et la question est de savoir comment en faire bénéficier le plus grand nombre.

Un autre aspect intéressant doit être pris en compte: les activités liées à l'information convergent dans un seul et unique monde virtuel, la toile mondiale, auquel on accède depuis son bureau (c'est-à-dire depuis son écran d'ordinateur). En d'autres termes, les technologies de l'information offrent des moyens radicalement différents d'enregistrer, de manipuler et d'utiliser l'information, mais elles exigent de nouvelles approches pour gérer et traiter cette information. Or, nous ne parvenons toujours pas à gérer l'information numérique de façon appropriée. La perte d'informations est l'un des risques des développements actuels, pas seulement en tant que telle, mais aussi parce que nous ne sommes pas conscients que ce risque existe. En conséquence, nous ne savons pas le détecter là où il faudrait. Il nous faut donc gérer l'information numérique et son accessibilité.

Les moyens traditionnels de recherche et d'accès à l'information ne suffisent plus dans un monde numérique, compte tenu de la nature différente de cette information. Contrairement aux documents classiques sur papier, les documents ou l'information numériques ne sont plus des entités matérielles fixes, et sont donc en tant que tels intangibles et volatils. Il nous faut donc de nouveaux mécanismes adaptés et complexes pour les rendre visibles, récupérables, accessibles et compréhensibles.

Plusieurs projets et approches s'efforcent actuellement de satisfaire à ces conditions. Certaines disciplines professionnelles, notamment les bibliothécaires et les archivistes, se préoccupent essentiellement de l'utilisation de métadonnées de toutes sortes, tandis que les constructeurs de logiciels proposent des outils tels que l'indexation automatique et la recherche en texte intégral. Les thésaurus et les normes constituent d'autres instruments. La question est alors de savoir quelle approche est la plus utile, le cas échéant, dans quelles circonstances et dans quel but?

Pour pouvoir utiliser des outils, il faut comprendre la finalité de leur utilisation — en l'occurrence: la fourniture de services électroniques, l'accès à l'information, la communication, la conservation, la gestion, etc. Toutes ces activités correspondront à des cahiers des charges différents, avec éventuellement certains recoupements partiels. Deux perspectives au moins peuvent être distinguées: celle de la création et de la maintenance de l'information et celle de son utilisation. Si le créateur et l'utilisateur ne sont pas la même personne, leur point de vue sera différent et, partant, leurs besoins le seront également.

Le web implique également d'utiliser un langage commun de recherche et de communication. L'interopérabilité de l'information des divers domaines et compétences devient donc de plus en plus pertinente. Hormis l'interopérabilité, la possibilité de récupérer les données n'est pas suffisante. Les informations récupérées et trouvées doivent également être présentées et testées quant à leur fiabilité, le dernier aspect — et non des moindres — étant la nécessité de les comprendre afin de pouvoir les interpréter. Tous ces différents aspects influeront sur la gestion de l'information numérique et sur son accessibilité.

De nombreuses initiatives sont déployées, qui s'efforcent de trouver des solutions pour rechercher l'information sur le web de façon pertinente. Ces initiatives sont dirigées pour la plupart par les besoins d'une communauté, comme dans le cas de l'initiative OAI (Open Archives Initiative) ou du projet DCMI (Dublin Core Metadata Initiative). Le développement du «web sémantique» en constitue un autre exemple récent.

En ce qui concerne les développements intervenant dans la communauté archivistique, l'encodage de la description (EAD) et l'encodage du contexte (EAC) des archives sont d'autres moyens de présenter des informations archivées à certaines communautés d'utilisateurs. Ces solutions reposent sur des séries de métadonnées d'archivage existantes.

L'exposé évoquera les aspects entourant la récupération et l'accessibilité des informations numériques, notamment la création de métadonnées, l'interopérabilité et les principaux aspects de l'évaluation des approches, des initiatives et des outils actuellement disponibles.