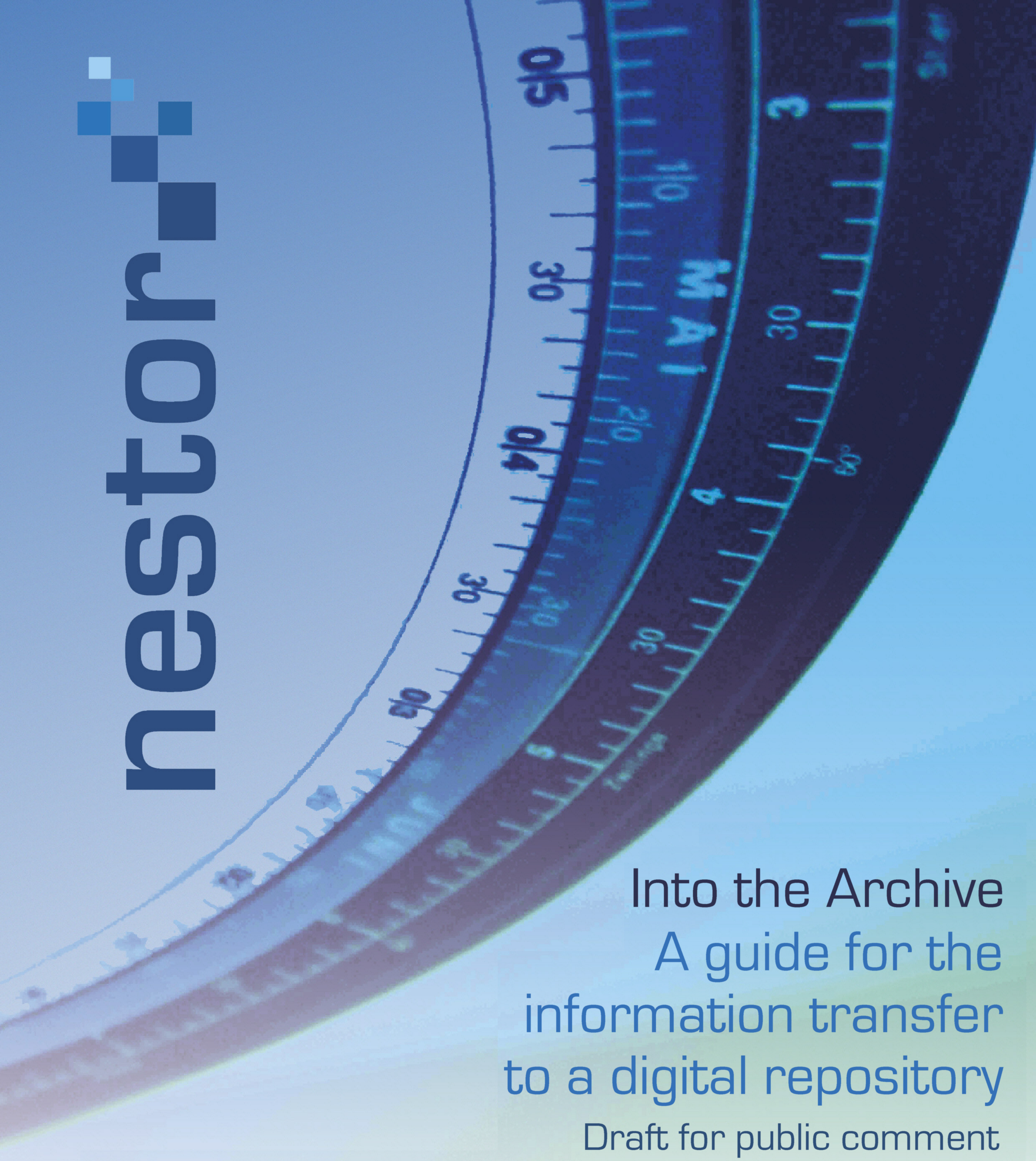

# Into the Archive

## A guide for the information transfer to a digital repository

Draft for public comment

nestor working group long-term preservation standards

nestor-studies 10

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

# Into the Archive

# A guide for the information transfer to a digital repository

Draft for public comment

nestor working group
long-term preservation standards

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

nestor-studies 10

published by

nestor - Network of Expertise in Long-Term Storage of Digital Resources

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

http://longtermpreservation.de
http://www.langzeitarchivierung.de

Project partners:

Bayerische Staatsbibliothek, München
Bundesarchiv
Deutsche Nationalbibliothek (Projektleitung)
FernUniversität in Hagen
Humboldt-Universität zu Berlin - Computer- und Medienservice / Universitätsbibliothek
Institut für Museumsforschung, Berlin
Niedersächsische Staats- und Universitätsbibliothek, Göttingen

nestor working group long-term preservation standards

c/o Niedersächsische Staats und Universitätsbibliothek Göttingen
Jens Ludwig
Papendiek 14
D-37073 Göttingen
Tel.: +49 (0)551-3912121
E-mail:<ludwig@sub.uni-goettingen.de>

URN: urn:nbn:de:0008-20080710002
http://nbn-resolving.de/urn:nbn:de:0008-20080710002

## Authors

Beinert, Tobias: Bayerische Staatsbibliothek
Büchler, Georg: KOST, Schweiz
Dr. Graf, Sabine: Niedersächsisches Landesarchiv
Huth, Karsten: Bundesarchiv
Dr. Keitel, Christian: Landesarchiv Baden-Württemberg
Ludwig, Jens: Niedersächsische Staats und Universitätsbibliothek
Rödig, Peter: Universität der Bundeswehr München
Steinke, Tobias: Deutsche Nationalbibliothek

## For their suggestions we are also grateful to:

Brandt, Olaf: SUB Göttingen/KB Niederlande, now BStU
Enders, Markus: SUB Göttingen, now British Library
Dr. Keller-Marxer, Peter: iKeep AG
Dr. Korb, Nikola: Deutsche Nationalbibliothek
Schrimpf, Sabine: Deutsche Nationalbibliothek
Dr. Wolf-Klostermann, Thomas: Bayerische Staatsbibliothek

Funded by the project Innovation with Norms and Standards (INS) 2008

INS
INNOVATION MIT NORMEN
UND STANDARDS

# Contents

## Introduction

The long-term preservation of digital data is an even younger topic of interest than digital data itself. And only now are a number of different projects in this field going into regular operation. As is commonly the case in new task areas, only a small number of standards and techniques have become established. "Best practices" as such have not yet emerged, as long-term preservation is only currently being practised by a handful of institutions. The aim of this guide is to help overcome the first obstacle encountered in the field of long-term preservation: how can digital information objects be ingested into a digital repository in a manner that facilitates their secure storage, management and preservation there?

The ingest of data into a digital repository is not merely a technical transfer between two systems, more importantly it is also a process which includes a large number of organisational requirements and which culminates in the acceptance of responsibility by the digital repository. However, a digital repository should not be regarded primarily as a technical system, but as an organisational unit or institution which performs the task of digital preservation and which is ideally based on the OAIS reference model. Many requirements which a digital repository needs to fulfil in ingesting information have already been formulated - in the nestor "Catalogue of Criteria for Trusted Digital Repositories" [Ref. 2], for instance. "Into the Archive" aims to clarify the goals and unique aspects of ingesting information into a digital repository: data needs to be transferred from usually heterogeneous and organisationally-specific contexts in such a way that it will nevertheless remain comprehensible and usable in completely different contexts in the future.

The ingest of information places demands not only upon the digital repository but also on the producer or supplier (possibly another digital repository) of the digital objects to be preserved. The special effort, complexity and relevance involved stem to a large extent from the fact that the digital repository and the producer both need to consider, in close collaboration, eventual usage scenarios and processes. May any changes be made to the material for legal reasons? Which of the material's properties definitely need to be preserved? And not least, the technical and other qualities of the materials have to be specifically checked, as these constitute the basis, which is frequently unalterable, for all subsequent measures.

Besides drawing upon the experience of those involved, the starting points for creating the guide were the OAIS reference model and the supplementary standard PAIMAS. However, these treat the ingest process as discussed here either primarily as a part of a digital repository or at a level of abstraction which is probably less useful as an introduction.

The designated communities for whom the guide has been created include:

- Memory institutions which, together with the information producers or the information-providing institutions, wish to realise a transfer and need a common basis upon which to initiate this.

- Information producers or information-providing institutions which wish to submit their information objects to a digital repository for long-term preservation and are looking for information on the forthcoming tasks.

- Memory institutions which are in possession of an infrastructure for long-term preservation and are now about to ingest information for the first time.

- Memory institutions which are planning a long-term preservation infrastructure and which are therefore addressing the issue of ingest.

# 1 Objects

## 1.1 Selection of information to be archived

The digital repository has to select the intellectual entities (see Glossary) to be ingested. For public institutions content-based selection of the information to be archived is generally derived from the institution's mission. The digital repository has to make the necessary technical decisions regarding the selection in collaboration with the producer. This primarily involves the selection of, and agreement upon, file formats suitable for long-term preservation which are required for the representation of information objects in the digital archive system, and also selection of the necessary metadata formats.

In some cases the standard export functions of the producer systems are not capable of exporting the information objects in the form desired by the digital repository. The producer system must then be extended by its supplier to meet the additional requirements, which is often very expensive. The entire ingest process is simpler and cheaper for both the producer and the archive if suitable export functions or archive interfaces are already taken into consideration and implemented when the electronic systems are being planned and procured. The digital repository should therefore be informed about the procurement of a new computer system from an early stage and be involved in its design and planning.

---

Objective:

*From the material provided by the producer, the digital repository should select the intellectual entities which are to be ingested on a permanent, unchanging and secure basis as archive objects. This selection is based on specific assessment criteria resulting from the legally or contractually defined mandate of a digital repository.*

Procedure:

Digital repositories collect information which can be read and interpreted by people, and which is stored in digital form. At this content level, it is permissible to use terminology derived from everyday experience in handling information objects (e.g. documents, files, films, photos, database of the Federal Statistics Office etc.). Naturally, defining the primary archive stock has consequences at the technical level.

An example of a standardised selection of intellectual entities is the DOMEA selection module [Ref. 4] which describes methods for the systematic submission of electronic files from government agency systems to the relevant archives and offers two methods for offering and selecting the appropriate files. In the first (4-stage) method, the archive is provided with an electronic list containing the possible objects, including the relevant metadata. The archive evaluates the desirability of archiving the files on the basis of this list. In the second method, the archive has direct access to the lists and to the files awaiting selection and can carry out its evaluation and selection there.

Objective:

*The digital repository and the producer should analyse the possibilities of the export interface of the producer system. If, after export, the data is not available in a form which the digital repository can make accessible on a permanent basis, suitable measures (e.g. conversion) are planned.*

Procedure:

Data by itself does not constitute an information object which can be read and interpreted by human beings. It contains coded information which needs to be correctly interpreted and represented by a suitable hardware/software environment. The technical equipment required for presentation of certain information objects can be very costly. This may be due to the costs involved in procuring the necessary technical equipment and maintaining it, or result from the limited lifespan of the equipment. If the digital repository sees no possibility of being able to guarantee representation of the information objects in the form offered by the export interface, the archive and producer will need to negotiate alternative formats for the information objects [Ref. 11].

Objective:

*There should be agreement between the archive and producer regarding necessary adaptations if the information objects cannot be archived in their existing form.*

Procedure:

The digital repository and the producer can be obliged to submit the information objects, to ingest them and to preserve them on a long-term basis. General rejection of ingest due to a lack of technical facilities on the part of the archive is not an option in this case.

The data is altered in case of a migration. The significant properties of the intellectual entity must remain intact, however, despite any changes (see Section 1.3). Selection of a data format which is capable of preserving the significant properties of the original format and which is also suitable for long-term preservation in the archive is one of the critical decisions in long-term digital preservation. The target format should be open and well-documented. Representation should be possible with significantly less technical equipment than that required by the original format [Ref. 11] [Ref. 15].

Migrating data to a different format represents an additional task which carries its own risks. The producer and the archive must clarify who is to carry which responsibilities and which costs. The selection of the technology required for carrying out migration can have a significant impact on the quality of the information objects and upon the the cost of the ingest. Conversion programs which purport to create the same target format do not necessary yield identical results. They can use completely different technologies (e.g. compression algorithms), all of which conform to the format definition, or deploy proprietary methods without expressly stating this. Emulation or supplements to the software/hardware environment of the digital repository are further possible options besides migration for presenting the ingested information objects.

## 1.2 Metadata selection

According to the OAIS standard, the ingest is not complete until a complete "archival information package" (AIP) has been created. The main constituents of an AIP are the content object itself plus metadata. Both are necessary to provide, on a permanent basis, sufficient information about the archive object (AIP), i.e. enabling it to be found, presented and comprehended and interpreted in context with other archive objects.

Crucial for long-term preservation is information which helps intellectual entities (e.g. magazine articles, files, photos etc.) to be created from information objects which can be interpreted by human beings. This information should contain references to the technical environments required for presentation, should identify the data format as clearly as possible (e.g. file format name and version) and should at least contain a reference to a comprehensive technical description of the file format (e.g. ISO standard, RFC, file format registry). In the case of complex information objects consisting of a number of files, the structure of the information object must be described in a comprehensible manner. Added to this is metadata which is necessary for the technical management of the data in the memory, such as file name, file size and hash values for validation of the archived data.

Metadata is also required which describes the content (e.g. author and title in the case of publications) of an archive object (AIP), and also the origin of an archive object (e.g. information stating which person from which authority submitted an electronic file to the archive at what time). A further important factor for reliable digital preservation is information on changes to the information objects which are made when they are exported from the producer's system. Just how detailed this information is represented in the metadata depends on the needs of the archive and the producer and the type of information objects being ingested.

Not all of the types of metadata described above exist immediately after an information object is exported from a producer's system. Some metadata is only generated during the ingest into the archive, and some must be requested from the producer. For this reason, before defining a transfer package, the archive must decide which information it requires both from the producer and from the producer's system.

---

Objective:

*The digital repository and the producer should define all metadata which is required on the selected information objects in the AIP. The result is a selection of all information needed to created a sufficiently comprehensive AIP.*

Procedure:

Information on the required technology and the structural links between the information object files is needed for the long-term, authentic representation of the information objects. Information to describe the semantic content and context of the information object is also necessary. A whole range of different types of metadata to be supplied by the producer should be considered. Besides the standard metadata used to describe the content, and the structural metadata already mentioned, this can also be technical and administration metadata.

Metadata formats can also be distinguished by content-based criteria, e.g. content-describing metadata (in the field of libraries: bibliographic metadata), technical metadata and legal metadata. There are specific standards for content-describing metadata such as MAB2 [Ref. 5] and MARC21 [Ref. 6] for the library sector and general standards such as Dublin Core [Ref. 7]. Dublin Core can be extended in the form of profiles, accordingly the partners must agree on whether only to use Dublin Core Simple or a specific profile. For technical and structural metadata there are PREMIS [Ref 3], METS [Ref. 8], LMER [Ref. 9].

The following are incorporated in the OAIS model [Ref. 1] for administration metadata:

- Provenance: Who did what, and when? The history of an object.

- Context: The relationships of the content outside the package. Why was it produced, what relations does it have to other content and packages?

- Reference: Identifier: Numeric or alphanumeric character strings which uniquely reference the intellectual entities and also the related information objects and identify them within the archive [Ref. 10].

- Fixity: Protection from unauthorised change, e.g. checksums (see also Validation section)

---

Objective:

*The digital repository and producer should reach an agreement on who will provide the necessary metadata.*

Procedure:

Not all the necessary metadata needs to be provided by the producer and submitted to the digital repository. It makes more sense if the description of the technology required for displaying an information object in the archive is provided by the archive itself. If the producer migrates the information objects to the PDF/A archive format before transferring them to the archive, this generates metadata which describes the producer's migration process and which should also be ingested into the archive. When the archive then checks that the submitted PDF/A files conform to the standard, this results in further metadata - this time in the archive - which can be integrated into the AIP.

---

## 1.3 Identification of significant properties of the intellectual entities and the information objects

In order to keep intellectual entities held in digital form accessible over longer periods of time, the information objects have to be represented in changing technical environments. The characteristics of the data are certain to change, regardless of whether the preservation strategy is based on emulation or migration. The significant properties are those characteristics which must remain constant under all circumstances.

---

Objective:

*The archive and producer should draw up a definition of the significant properties of the selected information objects.*

Procedure:

The archive and the producer have to decide which presentation deviations are acceptable and which are not. The needs of the archive users (designated community) should be a primary criterion. The InSPECT project lists the common subdivisions of the significant properties as: "

- content, eg. text, image, slides, etc.
- context, eg. who, when, why.
- appearance, eg. font and size, colour, layout, etc.
- structure, eg. embedded files, pagination, headings, etc.
- behaviour, eg. hypertext links, updating calculations, active links, etc. " [Ref. 16]

---

Objective:

*For each intellectual entity the archive ingests from the producer, it should record the relevant properties which should remain permanently intact in the metadata. This information can then be used later to check whether an upcoming migration or a new emulator is suitable for the long-term preservation of the objects.*

Procedure:

It is the responsibility of each individual archive to determine in which form the significant properties of an object are recorded. Hard technical values (e.g. image resolution, image width, colour spaces etc.) can be used, or the sensory impression of an intellectual entity can be described.

---

# 2 Processes

## 2.1 Definition of transfer packages

Once the digital repository and the producer have identified what is to be archived, it must be decided in which units the content is transferred to the digital repository: the transfer packages must be defined. This determines the relationship between the transferred data, the metadata and the information objects, thereby ensuring that each object can be reconstituted from its various parts. The transfer packages may be container formats which contain the constituent parts, or purely descriptive files which simply reference the data and metadata, thereby making it available to the archive. An example of a container transfer package for websites could be all files and metadata of a website being transferred together with a descriptive XML file in a ZIP file.

In the OAIS model [Ref. 1], such transfer packages are called SIPs (Submission Information Packages) and may differ from the differently structured outgoing packages (DIPs, Dissemination Information Packages) and internally used packages (AIPs, Archive Information Packages) of the archive itself. Transfer packages are also distinguished terminologically from package formats (which represent the technical format of the transfer packages) and from object models (which are used to convey the logical/conceptual properties of information objects).

It is important to specify the transfer package, as digital repositories and producers do not, in all likelihood, represent information objects in the same manner. A joint definition is therefore required of what constitutes a single unit to be transferred. Digital repositories often already have their own standards for transfer packages.

---

Objective:

*The relationship between an information object and one or more packages should be defined.*

Procedure:

Ideally, an information object will consist of one single package, as this reduces complexity. However, for technical or other reasons this 1:1 relationship is not often sensible or feasible, meaning that information objects often consist of a number of packages or that a package contains a number of information objects. For this reason a method needs to be defined (including corresponding metadata) which establishes the relationship between packages and information objects.

- It may be necessary to distribute large information objects across a number of different packages as the result of technical size limits.

- It may also be desirable to split an information object into a number of different packages if the submitted packages are similar to the eventual outgoing packages and only single access to individual parts of an information object is necessary for relevant usage scenarios.

- If certain data belongs to very large numbers of information objects, it may be more efficient to transfer a single package which is then referenced by many information objects and other packages. Format template files for large websites or document collections would be such examples.

---

Objective:

*It should be possible to reconstruct the data relationships on the basis of structural metadata contained in the package.*

Procedure:

Data is not initially related either to other data or to its metadata. What constitutes a relevant relationship for forming an information object depends on the technical environment and must therefore be made explicit.

Files can be held in a joint file directory or be named in accordance with a uniform system, expressing the technical and logical relationship between them. Where this is not sufficiently defined by formats, it may be necessary to explicitly describe the relationship between metadata and document files or dependencies of website files on a format template file. METS [Ref. 8], LMER [Ref. 9] and PREMIS [Ref. 3] e.g. provide appropriate options here.

---

Objective:

*It should be possible for the producer and the digital repository to identify the package.*

Procedure:

Identification can be provided by the identifier of the information object or associated archive package identifiers. Bear in mind, however, that there need not to be a 1:1 relationship between the packages and the information objects. Accordingly, each may need its own identifier. The identifier should be contained in the package; from a technical viewpoint, however, identification at the level of the transfer protocol is also conceivable. Persistent identifiers [Ref. 10] are preferable.

Archive packages can be identified by the URNs of the documents contained in them. All known metadata standards offer possibilities for this.

---

## 2.2 Validation

Given the ease with which it is possible to manipulate digital objects, a check should be made after the transfer to ensure that they still contain what they are supposed to. Validation is also part of the other phases of the archiving process. Basically, validation is necessary after any transfer - to a different facility, to a new format, to a new data carrier.

Validation is always a comparison. The purpose is to document the authenticity of the object (i.e. the object is what it claims to be) and its correct functioning. A distinction is made between two categories of comparison objectives associated with the object to be validated:

- The object to be validated is checked against its "parent" object (e.g. the hash values of the target file and the original file are compared after data carrier migration).

- The object to be validated is checked in terms of its formal or content specifications (e.g. a file format is compared with the description of the format).

It is possible to check a whole group of objects, and not just a single object. However, for reasons of simplicity, the generalised word "object" is used below.

---

Objective:

*Definition of the individual validation processes.*

Procedure:

The validation task can be subdivided into individual steps or processes. One process is a check to ensure that one or more characteristics of the ingested object has/have been retained. This can be carried out automatically or manually. The processes must be described and named, and clear dividing lines drawn up between them. For example:

- Does the delivery contain all the agreed objects?
- Are the objects intact (do they correspond to previously established hash values)?
- Are the objects free of all viruses?
- Are the files valid/error-free with regard to their file format?

---

Objective:

*The archive should define the required degree of compliance for each individual validation process, and the consequences of non-compliance.*

Procedure:

The purpose of some validation processes may simply be to determine fulfilment/non-fulfilment of a characteristic (e.g. hash values). In other cases, gradual transitions are possible. In this case the results often no longer meet the specified expectations in full (e.g. colour nuances). Sometimes it is not possible for the original and the target object to be identical (e.g. following migrations of the file format). It is even difficult to implement a standard (e.g. ISO 19005 - PDF/A) in full. The degree of fulfilment to be achieved must therefore be defined for each process. What must happen if this is not the case should also be defined. One consequence could be rejection of an object and its return to the producer (along with a request for submission of a flawless object). A further option would be only to record deviations up to a defined level in the archive metadata (in a validation report). PREMIS refers to these as "quirks" [Ref. 3, p. 204]. Possible examples of validation processes with different levels of fulfilment:

- n% defective entries in a database field are tolerated
- n% undocumented database characteristics can still be accepted

---

Objective:

*The individuals involved and the equipment used should be named for each validation process.*

Procedure:

It must first be clarified which tasks are undertaken by whom and which tools can be used for each process. For methodological reasons, in order to detect method and tool-related errors it may also be necessary to use other methods and software tools for validation than for generating the transfer package.

- Who carries out the validation: the producer or the archive?
- Are any third parties (experts, representatives of the designated community) involved?
- Which tools and methods are used for the validation?
- Where does the validation take place (in the producer's system or at the archive)?

---

Objective:

*The validation processes should follow a plausible chronological sequence.*

Procedure:

The validation processes are carried out in individual phases during the ingest. The division of the phases and the number of them can differ from archive to archive; mostly, however, there is one division into two phases. In the first phase, it is clarified whether the objects meet basic requirements immediately after they have been ingested into the archive. If there are any deviations from the expected result, ingest of the objects its rejected. These characteristics therefore have a disqualifying function. The more detailed validation processes are undertaken in the subsequent, second phase. Only here do the processes play a role which do not decide a clear yes/no result.

---

## 2.3 Transfer of data from the producer's system

Of great importance for the transfer of the data from the producer to the digital repository is the full and correct transmission of all data required by the repository to reconstruct the relevant information objects and for the long-term management, preservation and accessibility of the intellectual entities. Once the data has been transferred, the digital repository must be able to determine precisely who the delivery is from (authenticity), in which form the data and metadata need to be submitted (validity) and how large the data deliveries should be (completeness). The legal and contractual conditions are decisive here, and these may vary from case to case. The legal requirements and the technology used for the transfer must be harmonised.

Objective:

*The legal and/or contractual framework for a transfer should be analysed and defined by the producer and the archive.*

Procedure:

The requirements for a transfer with regard to aspects such as secure data transfer, ongoing validity of qualified signatures etc. [Ref. 12] [Ref. 13] may vary. Files delivered from a government agency, for instance, may have different confidentiality levels.

Objective:

*The producer and the archive should be aware of the technical and organisational possibilities available for the transfer. Both should know whether the technical facilities permit transfer which conforms to the requirements or not. If necessary, agreement should be reached concerning the necessary adjustments.*

Procedure:

Both sides must know their technical capabilities and carry out checks, especially in cases where no transfer has yet been made between the partners. Before the first transfer in particular, the producer's and the archive's technical possibilities may not conform to the legal requirements. Conformity should be carefully checked, as IT security conditions are also subject to permanent change. For example, encryption and qualified signature protocols lose their effectiveness against malicious manipulation over time.

Objective:

*The individual work stages of the transfer between the producer and the archive should be precisely coordinated and tested.*

Procedure:

The transfer process is critical for the authenticity of the information object. Exact adherence to an agreed transfer process increases the trustworthiness of the archive. In the case of regular, automatic transfers, the transfer process needs to be implemented technically in the systems of both the producer and the archive. The producer and archive stipulate the maximum file size for the transfer, agree on the technical reports which need to be created, the transfer method (data transfer with protocol, or on data carrier with data carrier format information provided by means of delivery), the time period of the transfer, all necessary identifiers and passwords and the necessary security protocols.

The agreed technical tools need to be implemented in the systems of the partner institutions and coordinated. For safety's sake, the producer and the digital repository jointly generate test transfer packages and conduct controlled transfer tests. Only under controlled conditions can it be determined whether the agreed and implemented transfer is functioning correctly.

# 3 Management

## 3.1 Identification of legal and contractual conditions

Statutary regulations between the producer and the archive must exist, or be created, before any digital objects are ingested into the archive, in order to ensure long-term planning and legal security for both sides. Legal questions arise here, which are not directly related to the ingest but which nevertheless need to be clarified, on the ingest of an object in order to regulate the permanence of the archiving and the conditions for handling the archived objects. Clarification of further legal issues affecting the producer-archive relationship is therefore a precondition for effective and successful ingest. After analysis of the basis of the legal relationships between the producer and the archive, copyright issues are the next main focus in this field.

Objective:

*All parties acting in a legal capacity, and the persons authorised to represent them, should be identified or appointed.*

Procedure:

The task of archiving must be established on a legal basis in order to provide legal and planning security both for the producer and also for the archive. It should be clarified whether such statutary regulations exist. If such a legal basis exists, its nature must be determined. It should be established whether the archiving activity is, or will be, based upon a statutary mandate or upon a legal agreement between the archive and the producer. If there is no legal basis in the form of a statutory deposit obligation, an agreement (licence agreement) needs to be created between the producer and the archive, at least concerning copyright issues. For the archive to carry out its work effectively it must be given the necessary rights for the planned archiving and usage in the form of a statutory regulation or a contractual agreement.

Objective:

*The obligations of the archive and/or producer with regard to handling the material to be archived should be known.*

Procedure:

It should be determined whether binding requirements regarding the storage and use of the objects to be archived, or their content, derive from the statutory regulations.

An archive law could, for example, specify how many copies of an object should be stored in the archive and whether the producer should delete original documents. Also, it is possible that data protection requirements may prevent a comprehensive search of the archive stocks or the provision of archive material to third parties.

Possible legally regulated responsibilities on the part of the producer could include his being obliged to offer or deliver the digital objects to the archive. The distribution of the costs must also be legally regulated, including transportation to the archive, the costs for archiving and care, and the costs for generating copies on behalf of the producer.

---

Objective:

*The archive should be aware of the copyright conditions attached to the material to be archived and permanently record them.*

Procedure:

The rights of the copyright owner of the object to be archived must be clarified. Under certain circumstances the material to be archived may be subject to intellectual properties rights. If the relevant regulations apply, digital archiving is only permissible in Germany in highly restricted circumstances, as the archiving of digital objects always includes duplication as defined in the Copyright Act (UrhG). Such duplication must always be covered by a corresponding regulation of the Copyright Act or by transfering the relevant usage rights from the rights owner to the digital repository.

In Germany the provisions of the Copyright Act (UrhG) only apply if a work has a certain threshold of originality and it falls within the legally defined period of protection. A regulation of property is contained in Art. 53 of the German Copyright Act, the so-called archive regulation, which permits duplication for the purpose of transfer to an organisation's own archive. However, the archive regulation only permits duplication for the purposes of collection, storage and conservation, but not for use of the archived objects by third parties.

---

Objective:

*The producer and archive should have analysed the content of the copyright requirements and, if necessary, have established appropriate regulations.*

Procedure:

If copyright issues are attached to the digital objects to be archived, various long-term preservation problems need to be taken into account which either need to be regulated on a legal basis or require a legally valid agreement between the established owner of the relevant rights and the archive. To enable the archive to carry out its work it has to be given the necessary rights for the planned forms of archiving and usage in the form of a statutory regulation or a contractual agreement. The responsibilities of the producer and archive need to be precisely defined on a statutory basis and / or by means of binding agreements.

The migration of digital objects to other file formats, for instance, could be permitted by a special statutory regulation. The removal of digital rights management (DRM) needs to be contractually regulated between the producer and the archive if there is no statutory basis for this. The archive and the producer/rights owner can specify in a licence agreement that the archived objects are to be made accessible to a specified group of users.

Objective:

*Warranty and liability issues should be regulated between the producer and the archive.*

Procedure:

It should be laid down when damage claims can be pressed, and by which party, and which duties of care need to be observed. It should also be clarified whether regulations exist which apply when the rights of third parties are violated by the producer or the archive.

## 3.2 Ingest agreements and documentation

The documentation of the agreed ingest standards and specifications, and of the reporting of the ingest procedures, lend transparency to a part of the provenance of the objects to be archived. In this way they help ensure the integrity and authenticity of these objects, as required by the criteria catalogue of trusted repositories. The same requirements apply to archiving as to the primary data.

Objective:

*The producer and the archive should draw up an ingest agreement. This is a binding agreement which regulates all aspects of the information ingest.*

Procedure:

The agreement must be approved by the producer and the digital repository. It records the results of the ingest planning as described in detail in this document. The agreement serves as a binding manual for the actual ingest process.

Any changes to the process elements laid down in the agreement must be made in a regulated and documented procedure. In particular, the agreement should contain:

- the list of the intellectual entities to be archived, including a definition of their significant properties;
- the list of information objects and data representing these intellectual entities; the technical environment required for archiving them and any migration agreements;
- the list of required metadata, stating who has supplied it;
- the transfer package format including the required metadata in it, the identifier and the assignment of the information objects into packages;
- the transfer and its technical implementation;
- the definition of the individual validation processes including their required degree of fulfilment, the consequences in the event of non-fulfilment, the persons and tools involved and the chronological sequence;
- the information which serves as the basis of the risk analysis; especially the estimates of the data quantity, computer capacity and computing time required for ingest and cost estimates;
- the parties acting in a legal capacity, their relationships and the regulations regarding copyright and liability;
- the schedule for carrying out the ingest.

Objective:

*A report should be made on each ingest, from the start right through to archiving. This protocol should be preserved for the same duration as the objects themselves in the archive.*

Procedure:

The protocol could contain the following information, for example:

- List of all intellectual entities ingested, plus corresponding information objects
- Name of producer
- Time and date of the start of the transfer.
- Time and date of the arrival of the transfer package at the archive.
- Time and date of archiving.
- Transformations undertaken on information objects.
- Results of individual validations.

## 3.3 Quality, security, process and risk management

Objectives of management activities: The ingest of digital information is a critical process which requires appropriate relevant management activities to achieve the required level of quality and to deal with risks appropriately with regard to security and costs. The management should have an overview of the entire system and ensure that it is effective, i.e. that the targets are realistic and have been defined in line with the specifications of the sponsoring body and the legislator, that a suitable organisational structure and infrastructure have been set up for achieving the targets and that all processes have been harmonised. A range of different but compatible management systems need to be established for this. The aim here is to avoid monitoring the ingest in isolation, which results in unnecessary fluctuations in the quality, security and local efficiency and also in unanticipated risks.

Quality management: The primary target of a quality management system is customer satisfaction, but also the satisfaction of other interested parties and public service providers or society as a whole. The main tasks of a quality management system are to recognise the customers' current and future requirements and to implement their demands. An organisation needs to lay down a quality policy and quality targets for this. The processes and responsibilities required to achieve the quality targets need to be defined. The required resources should also be determined and provided (cf. process management). The effectiveness and efficiency of the processes should be checked and the causes of errors analysed and remedied. Within the context of the ingest, quality management ensures e.g.:

- basic standards for all ingest and validation processes in compliance with the quality targets of the organisation.
- provision of all necessary resources for quality management processes including validation.
- high quality of the necessary resources.
- revision of the quality policy and quality targets if these are not achievable.
- standardised check of all validation processes.
- basic standards for the documentation of all ingest processes.
- integration of any outsourced processes into quality management.
- identification and application of suitable quality standards.

Information security management: The objective of security management is to ward off threats which put achievement of the overall targets and ultimately also the trustworthiness of the organisation at risk. The task of security management is to assume overall responsibility - in this case for information security. A security policy and security targets need to be laid down for this. The relevant processes and responsibilities should be determined for implementation, and the required resources made available. Of prime importance are recognition of threats and calculation of the risk potential. The effectiveness and efficiency of the security processes should also be checked and the causes of errors analysed and remedied. Within the context of the ingest process, information security management provides e.g.:

- security policy and security targets which have been coordinated with producers and suppliers and laid down in agreements.
- standards which take into account the specific risks of categories of ingests (self-archiving of a community, anonymous access, ingest of executable objects, classified materials or virtual objects etc.).
- appropriate organisational conditions (e.g. appointment of a security officer, definition of responsibilities for issuing passwords).
- the (possibly joint) provision, operation and monitoring of a suitable security infrastructure (e.g. public key infrastructure, definition of suitable metadata).
- identification and application of security standards or statutory requirements.

Process management: The target of process management is the effective and efficient implementation of organisational targets. It ensures internal and external transparency, thereby making a contribution to the effectiveness and trustworthiness of an organisation. Process management tasks include summarising all process activities, taking temporal, spatial and sequence-logic dependencies into account, thereby enabling economically meaningful statements to be made. This task is based on the organisation's targets. Processes form the basis for allocating resources and responsibilities and for specifying the required performance in each case. This permits organisational and technical interfaces to be defined more precisely. Process management tasks also include the integration of different management processes.

Within the context of the ingest, process management ensures e.g.:

- highlighting of all ingest effects on subsequent processes, especially on archival storage and the provision of information.

- the avoidance of irregular effects on other processes through measures aimed at quality assurance, security or an increase in the efficiency of ingests.

- consideration of the processes preceding the ingest. Knowledge of this or, better still, the ability to influence it, can have a positive impact on the ingest (selection of formats for content and metadata, knowledge about technical and intellectual creation contexts).

- the inclusion of ingests carried out or prepared on a project basis into "standard processes" by means of appropriate configuration and change management.

- adjustment of existing ingest processes to altered technical and organisational conditions on both the archive and customer side through appropriate configuration and change management.

Risk management: Besides the risks to quality and security, additional risks may arise depending on the way in which ingests are organised. The aim of risk management is to record and assess the risks, to minimise, control and monitor them on an ongoing basis. Five different risk management strategies are distinguished between: avoiding, reducing, limiting, shifting and accepting risks. Risk management is an iterative process and covers the entire ingest process. It involves both the producer and the archive. Within the context of the ingest, the following risks should be considered:

- Financial risks: Every ingest process (and every aspect of the ingest process described in this manual) demands resources from the producer and the archive and generates a need for further resources for the permanent archiving of the ingested data. The detailed planning and the budget of an ingest process serve as the basis for the management of financial risks. Financial risks affect both human and financial resources.

  - Human resources: The labour of a range of staff with different skills is required to carry out the ingest process. Permanent archiving of the ingested data then necessitates further human resources for any required migration, and for processing the data for use.

  - Financial resources: The ingest process generates costs for the data transfer, provisional storage and the required computer capacity. Permanent archiving of the ingested data generates costs for memory storage, data backup and data protection, for computer capacity and temporary memory for use of the data, and computing time and temporary memory for any migrations.

- Legal risks: The ingest of information by an archive takes place within the framework of various legal and/or contractual provisions (as set out in "Identification of legal and contractual conditions"). Non-observance of these provisions can result in a range of sanctions.

- Reputation risks. Besides legal considerations, an archive may also be subject to certain moral obligations regarding the long-term storage of digital data. Accordingly, the reputation of the archive can suffer by not ingesting such data, or by ingesting the data but not ensuring its preservation.

# Glossary

**Archive:**

See Digital repository.

**Authenticity:**

The object is what it claims to be. (taken from: Criteria catalogue of trusted repositories, Ref. [2] p. 34)

**Data:**

Digitally stored elements of an information object (according to PREMIS: files or bitstreams). Corresponds to the PREMIS object type "file" and/or "bitstream". (cf. PREMIS 2.0, Ref. [3] p. 7)

**Designated community:**

An identifiable group of potential users with specific interests and circumstances. It could be the general public or a group of specialist scientists, for instance. It can be heterogeneous and consist of different user groups. (taken from: Criteria catalogue of trusted repositories, Ref. [2] p. 35)

**Digital repository:**

An organisation (consisting of people and technical systems) which has assumed responsibility for the long-term preservation and long-term availability of digital data and its provision for a specified designated community. "Long-term" here means lasting beyond technological changes (to hard and software) and also any changes to the designated community (e.g. for future generations, indefinitely). "Archive" is used in the text as a synonym for digital repository. (taken from: Criteria catalogue of trusted repositories, Ref. [2] p. 34)

**Emulation:**

Strategy for preserving the long-term accessibility of digital objects. The strategy ensures that the system requirements for using older digital objects can be recreated (emulated) through the use of special software on common systems currently available on the market. The digital objects themselves remain unchanged where possible.

**Information object:**

Consists of digitally stored data units (according to PREMIS - files or bistreams) and can be an intellectual entity. Corresponds to the PREMIS object type "representation". (cf. PREMIS 2.0, Ref. [3] p. 7)

**Ingest:**

Signifies the organisation and execution of all processes necessary to accept an information object into the archive and for the archive to assume responsibility for it.

**Integrity:**

[1.] Completeness of the digital objects [2.] and exclusion of unintended modifications as defined in the preservation rules. The yardsticks for integrity are the characteristics of a digital object which are defined as worthy of preservation. (taken from: Criteria catalogue of trusted repositories, Ref. [2] p. 34)

**Intellectual entity:**

Logically discrete unit of content which can be interpreted by human beings and can be represented materially/physically by information objects. (cf. PREMIS 2.0, Ref. [3] p. 6)

**Long-term preservation:**

The long-term preservation of digital objects includes all measures aimed at preserving digital objects permanently for future generations. The term is closely related to long-term accessibility, although the emphasis of the latter is on the permanent usability of the data. Common long-term preservation strategies include emulation and migration. (taken from nestor glossary)

**Metadata:**

Data representing information about other data by describing e.g. its content, structure, composition, handling, origin etc. (taken from: Criteria catalogue of trusted repositories, Ref. [2] p. 34)

**Migration:**

File format migration: Conversion of an information object from one data format into another. A preservation measure to adapt a digital object to a changed technical environment. Data carrier migration: Copying an information object to a different data carrier

**Producer:**

People or client systems who/which transfer digital objects to the digital repository for long-term preservation. They are not necessarily the originators; they could also be the suppliers of the digital objects. (from Catalogue of Criteria for Trusted Repositories, Ref. [2] p. 35)

**Significant properties:**

Characteristics of an information object which are considered as important for the designated community and which should therefore be preserved. (cf. PREMIS 2.0, Ref. [3] p. 39)

**Transfer package:**

Defined number of information objects transferred as a unit by a producer to the preservation repository, for example as one file or a group of files.

# Overview of the targets of information ingest

1. **Objects**
    a. Select information to be archived
        i. Select intellectual entities
        ii. Analyse export form
        iii. Agree necessary adaptations
    b. Select metadata
        i. Define required metadata
        ii. Clarify responsibility for providing the metadata
    c. Significant properties
        i. Define the significant properties
        ii. Compile the significant properties

2. **Processes**
    a. Transfer packages
        i. Assignment of the information objects to transfer packages
        ii. Reconstructability of information objects
        iii. Identification of the transfer packages
    b. Validation
        i. Definition of validations
        ii. Required degree of fulfilment and consequences of non-fulfilment
        iii. Persons involved and tools
        iv. Schedule
    c. Transfer of data
        i. Legal/contractual framework
        ii. Technical and organisational possibilities
        iii. Definition and test of transfer work stages

3. **Management**
    a. Laws and contracts
        i. Identification of corporate bodies under public law and agents
        ii. Definition of relations between producer and archive
        iii. Obligations concerning archive materials should be known
        iv. Copyright ascertained
        v. Regulation of copyright
        vi. Warranty and liability
    b. Ingest agreement and document
        i. Binding documentation of decisions
        ii. Reporting of ingest processes
    c. Areas of management
        i. Quality
        ii. Safety
        iii. Processes
        iv. Costs and risks (legal, reputation, personnel, finances)

# References

**Basics:**

**[1]** ISO 14721:2003 - Reference Model for an Open Archival Information System (OAIS) – Space Data and Information Transfer Systems

**[2]** Catalogue of Criteria for Trusted Digital Repositories/ published by nestor Working Group on Trusted Repositories Certification. Frankfurt am Main : nestor c/o Deutsche Nationalbibliothek, June 2006. - nestor Materialien 8 – URN: urn:nbn:de:0008-2006060703

**[3]** PREMIS Data Dictionary for Preservation Metadata – version 2.0 – PREMIS Editorial Committee, März 2008

**Referenced literature and standards:**

**[4]** DOMEA Konzept – Organisationskonzept 2.0 – Erweiterungsmodul zum Organisationskonzept 2.0 Aussonderung und Archivierung elektronischer Akten – Schriftenreihe der KBSt 66, Oktober 2004 – Website: <http://www.verwaltung-innovativ.de/cln_110/nn_1007474/SharedDocs/Publikationen/DE/domea__konzept__aussonderung__und__archivierung__elektronischer__akten.html?__nnn=true >

**[5]** MAB2: Maschinelles Austauschformat für Bibliotheken. - Loseblatt-Ausg. - ISSN 0949-5258 - Grundwerk . - 2. Aufl. (auf dem Stand der 1. Erg.-Lfg. Mai 2002). - ISBN: 978-3-933641-00-7

**[6]** MARC Standards – Library of Congress – Network Development and MARC Standards Office – Website: <http://www.loc.gov/marc/>

**[7]** The Dublin Core Metatdata Element Set ISO 15836-2003 - Dublin Core Metadata Initiative – Website: <http://dublincore.org/>

**[8]** METS: Metadata Encoding and Transmition Standard – Library of Congress – Website: <http://www.loc.gov/standards/mets/>

**[9]** LMER: Langzeitarchivierungsmetadaten für elektronische Ressourcen – Version 1.2 – April 2005 – URN: urn:nbn:de:1111-2005041102

**[10]** Persistent Identifier …eindeutige Bezeichner für digitale Inhalte… - Website: <http://www.persistent-identifier.de/?link=900>

**[11]** SAGA: Standards und Architekturen für E-Government Anwendungen – Version 4.0 – Bundesministerium des Inneren, März 2008 – Website: <http://www.kbst.bund.de/cln_012/nn_837392/SharedDocs/Meldungen/2008/SAGA/saga__4__0.html>

**[12]** BSI Standards – Bundesamt für Informationssicherheit – Website: <http://www.bsi.de/literat/bsi_standard/index.htm>

**[13]** ArchiSafe – Physikalisch-Technische Bundesanstalt – Website: <http://www.archisafe.de/s/archisafe/index>

**[14]** ISO 20652:2006 - Producer-Archive Interface Methodology Abstract Standard (PAIMAS) - Space Data and Information Transfer Systems

**[15]** ISO-19005-1 - Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)

**[16]** Andrew Wilson, Significant Properties Report, 2007. Website: <http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf >