Technology Watch Report

File formats for preservation

Malcolm Todd
The National Archives

DPC Technology Watch Series Report 09-02
October 2009

Executive Summary:

File formats are the principal means of encoding information content in any computing environment. Preserving intellectual content requires a firm grasp of the file formats used to create, store and disseminate it, and ensuring that they remain fit for purpose. There are several significant pronouncements on preservation file formats in the literature. These have generally emanated from either preservation institutions or research projects and usually take one of three approaches:

- recommendations for submitting material to digital repositories
- recommendations or policies for long term preservation or
- proposals, plans for and technical documentation of existing registries to store attributes of formats.

More recently, attention has broadened to pay specific attention to the significant properties of the intellectual objects that are the subject of preservation.

This Technology Watch Report has been written to provide an overview of these developments in context by comparative review and analysis to assist repository managers and the preservation community more widely. It aims to provide a guide and critique to the current literature, and place it in the context of a wider professional knowledge and research base.

At the time of writing, there is apparent consensus on five main criteria for file format selection:

- adoption: the extent to which use of a format is widespread
- technological dependencies: whether a format depends on other technologies
- disclosure: whether file format specifications are in the public domain
- transparency: how readily a file can be identified and its contents checked
- metadata support: whether metadata is provided within the format

There are other commonly-expressed criteria, such as:

- reusability / interoperability: can the format function with a variety of services
- robustness / complexity / viability: is the format inherently simple
- stability: is the format part of a managed release cycle and is this
- intellectual property / digital rights protection: whether rights complicate preservation

The main finding of this report is to support the proposal by Rog and van Wijk of the National Library of the Netherlands (2008) that such criteria should be used as a tool to work out the detailed implementation of a clear preservation strategy according to a prioritisation *appropriate to the repository*. This is essential to make sense of an otherwise bewildering array of considerations and provides key governance to ensure a preservation institution is managing the risk of obsolescence to its holdings. Such an approach is both more useful and more realistic than attempting a "definitive" list of formats or even selection criteria: both will vary with the circumstances of a repository.

The report also draws attention to three criteria that the preservation community addresses only sporadically in the file format literature:

- the ability of formats to convey content information
- extent, or 'verbosity' of format and
- cost

The first is examined from an archival perspective. Metrics for the second are clearly part of the bread-and-butter of many repository managers' current activities but it is surprising that it rarely appears in this context. The third requires further enquiry beyond the scope of this report.

In addition it would appear that much of the literature on formats has become remote from the terms used in the OAIS Reference Model - the key reference for the preservation community. Although the terminology and level of OAIS does not lend itself to discussing the detail of these issues it contains key concepts that need to be mapped consistently. The ancillary finding is that terminological looseness is not helping either practice or research.

Given the exciting stage the research has got to in defining significant properties of intellectual objects and the technological environments they are now being created in, this needs urgently to be addressed by the preservation community. The present review of the ISO-standardised incarnation of OAIS, ISO14721 gives an excellent opportunity to do this. There are several parts of our own community using different terms (or, worse, the same terms to mean different things). We also need to facilitate exchange with other disciplines especially our producers, users, policy-makers and funders.

Keywords:
Archives, file formats, OAIS, Open Archival Information System, migration

About the author

Malcolm Todd has been involved in managing and preserving digital records for over a decade. He was in records management practice in a government agency until moving to the Public Record Office (now The National Archives) in 2001 to specialise in digital records. Malcolm wrote significant parts of the guidance to support digital records projects in UK Government as well as managing the prominent TNA software testing scheme for ERMS between 2002 and 2004. From 2006-08 he was on a secondment to the UK Parliament working on digital preservation, metadata and ERMS and pilots. Whilst there he chaired the European Commission's review group drawn from the DLM Forum to advise on the ongoing development of MoReq2. He is now back as part of Archives Sector Development at TNA, leading on digital preservation outreach to public institutions beyond TNA.

Malcolm has been involved in a number of international collaborative initiatives. He has frequently been part of the UK delegation to ISO TC46/SC11 (Archives and records management). From 2003-05, he was on the Advisory Board of the Clever Recordkeeping Metadata Project at Monash University, Melbourne. From 2004-06 he was the European co-chair of the Policy Group within the InterPARES2 Project (University of British Columbia) investigating the preservation of authentic records from dynamic, interactive and experiential systems in the arts, sciences and government. He is a member of the editorial review panels of the Records Management Journal and the Digital Curation Manual (Digital Curation Centre). He has published papers with *Archival Science* (Springer), *Archivaria* (Association of Canadian Archivists) and the Society of American Archivists. This is his first formal collaboration with the Digital Preservation Coalition.

Acknowledgements

The author and DPC gratefully acknowledge David Giaretta and the Consultative Committee for Space Data Systems for permitting reproduction of Figures 1 and 2, and Gareth Knight of Centre for E-Research (CeRch) at King's College London for permitting reproduction of Figure 3.

Table of Contents

1 Introduction 6
1.1 How to read this report 6
1.2 Background, scope and audience 6
1.3 Definitions and terminology 6
1.4 OAIS perspectives 7
1.5 A theoretical "minimal redundancy" paradigm 9
2 Recommendations for action 10
2.1 For repository managers 10
2.2 For the wider preservation community 10
3 Current file format recommendations 13
3.1 Methodology 13
3.2 Normalising the discussion and the extent of consensus 13
3.2.1 Rearticulating the "Standards debate" in terms of adoption and disclosure criteria 14
3.3 Outside the core criteria 15
3.4 The 'absent' criteria: cost, extent and ability to represent full content 15
3.5 Grouping, weighting and hierarchy of criteria 16
4 Reconciling contrary criteria and scores 18
4.1 Divergent criteria 18
4.2 Role of preservation strategies 19
4.3 Revisiting some prominent digital preservation strategies 20
4.3.1 National Library of the Netherlands 20
4.3.2 Public Record Office of Victoria 20
4.3.3 National Archives of Australia 20
5 Preservation tools and infrastructure to support strategy implementation 22
5.1 File format and representation registries 22
5.2 PREMIS metadata dictionary 22
5.3 Current European research outputs in characterisation and preservation planning 22
5.4 Significant properties 23
6 Contributions from archival research and practice 25
6.1 The 'performance' model, ISO 15489 and content vs. documentary form 25
6.2 InSPECT Project: Canonicising significant properties of simple digital records 26
6.3 Archival science and the InterPARES2 Project: resolving the content and authenticity problems in complex dynamic, interactive and experiential environments 27
6.3.1 Broader application of archival viewpoints within digital preservation community 29
6.3.2 Extrapolating consequences of archival viewpoints into preservation file formats for distributed and web computing 30
7 Defining significant properties: challenge of parsing representation and preservation description information 32
8 Conclusions and recommendations 33
9 References 34
10 Table of core and wider criteria 38
11 Glossary 40

1. Introduction

1.1 How to read this report

This report is structured to permit rapid assessment of recommendations by operational managers and more detailed scrutiny in the context of archival and information sciences through emerging theory and practice in long term data management. It is structured differently from other reports in this series. The recommendations are presented immediately after a discursive introduction. There then follows a more detailed discussion of how format selection criteria have been derived and how they can be reconciled, the tools available to support strategy implementation, and a review of the implications from archival science that also discusses the concept of 'significant properties'.

1.2 Background, scope and audience

The selection of appropriate file formats is likely to be a significant factor in the survival of information objects. Digital information depends on hardware and software to make it comprehensible, so changes in hardware and software mean that long-term access requires a degree of coordination between original data and current facilities. There are different ways to create this coordination to make data accessible, but all approaches assume a firm grasp of the conventions used to encode it in the first place.

This report summarizes trends and issues in file format selection as they are articulated in the digital preservation literature at the time of writing (2008-9). This reflects the current status of digital preservation implementation: some of it concerns current practice from functioning repositories while more developmental approaches continue to emanate from the research community. Pronouncements by operational archives are more frequently in the form of their submission guidelines than a full statement of their ongoing preservation formats, though the two are related. As a result, whilst this report concentrates on formats of files being managed by archives (Archival Information Packages in OAIS terminology), it also considers literature about preferred submission formats (i.e. within Submission Information Packages or SIPs). Dissemination formats (Dissemination Information Packages or DIPs) are given less consideration.

The report considers file formats in the context of other elements of a digital preservation infrastructure, especially representation registries, archive governance and digital preservation strategies. There is interdependency between these issues. Some of the current literature appears to minimise this fact by treating format migration or 'conversion' almost in isolation[1]. To understand why this might be, it is helpful to review the Open Archival Information System information model and particularly its terminology. OAIS does not generally concern itself with the issues discussed in much of this report, but this does not mean that the deeper implications of the OAIS information model do not need to be taken into account when considering file format issues.

1.3 Definitions and terminology

As with many digital preservation reports, readers might want to have some familiarity with OAIS terms. This report uses OAIS terms where they are convenient and reasonably accessible. In this section we consider why OAIS issues rarely surface in the literature on preservation file formats. A table of other terms is included in Annex 2 to help readers. This includes definitions commonly used within

[1] "Conversion" is the term used in the records management standard ISO15489. At the time of writing a new work item on "digital records conversion" has been accepted from ARMA International by the relevant ISO technical committee, with an immediate need to clarify this usage

the digital preservation community and used here, along with a number of other terms from the records, library and particularly the archival science communities. Typically, terminology seems to hamper the consistent articulation of file format issues and could cause problems for research and practice in the future. This is particularly the case as collaboration moves into the area of significant properties where precision is paramount. Some of these issues are discussed as they arise in the text, but are clearly flagged or referenced to their source in the literature (several of this report's recommendations arise in this area).

Even the concept of "file format" is not completely settled: as noted by McLellan (2007 1-3). She draws distinctions between tagged textual formats, wrapper formats and what she describes as 'true' file formats (something she only seems to define in terms of being different from the "wrapper" and "tagged" categories). The issues she raises are valid but handled slightly differently in this report (see transparency in the core criteria and distributed and web computing).

The following three definitions are reasonably consistent as to concept:

> *a class of digitally-encoded assets defined by a set of semantic, syntactic and serialisation encoding rules for converting from abstract information to tangible byte streams (Abrams 2007, 51);*
>
> *a specific, pre-established structure for the organization of a digital file or byte-stream (PREMIS 2008, 195);*
>
> *the organization of data within files, usually designed to facilitate the storage, retrieval, processing, presentation, and/or transmission of the data by software (InterPARES2 n.d.).*

The first is the preferred definition in this report. All three have subtle differences to working definitions from the Library of Congress:

> *digital content formats that are independent of the physical medium on which they are stored or transported. Content in such formats exists as data files or data streams* (Arms and Fleischhauer 2005, 1);
>
> *packages of information that can be stored as data files or sent via network as data streams (aka bitstreams, byte streams[2]).*

The main difference is that a file format in the first three must comply with a defined specification above the symbolic encoding level, *recognized by relevant software and operating systems as a discrete and finite digital object*. This is an important issue in identification, representation, validation, rendition and management.

1.4 OAIS perspectives

A brief annotation of the key part of the OAIS information model follows:

[2] http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml,at the same time as quoting a working definition from GDFR. The latter, unsurprisingly, is closer to Abrams in this respect.

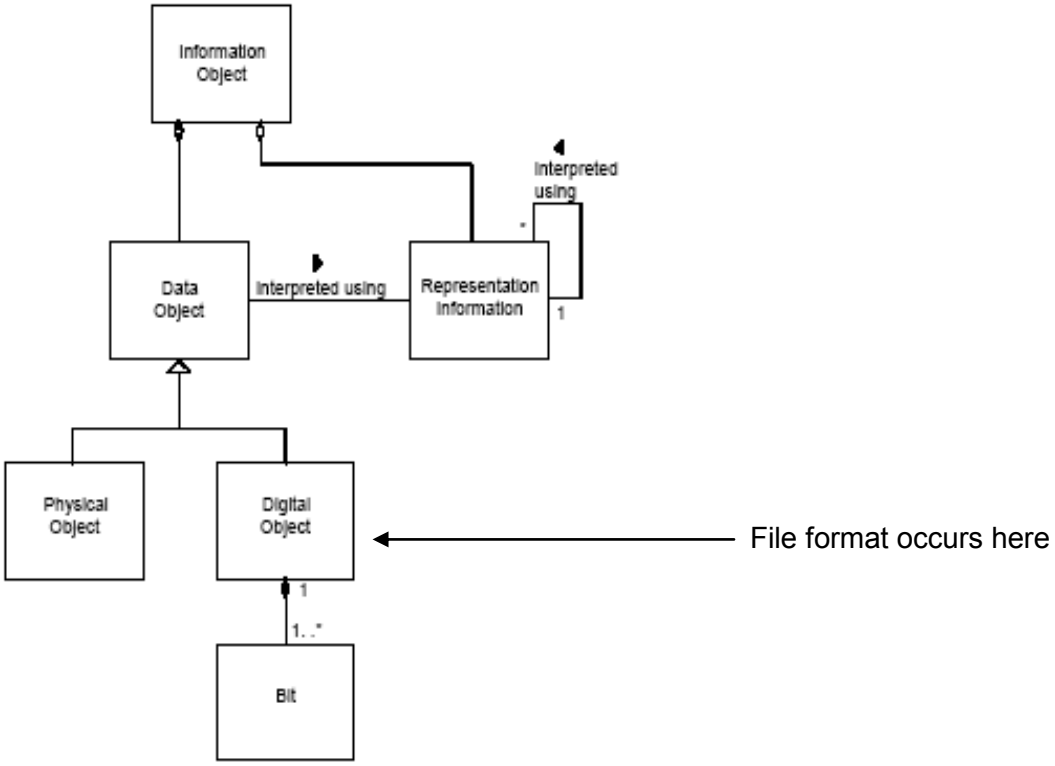

Figure 4-10: Information Object

Figure 1: OAIS information object model (lower levels) - (figure 4-10 within CCSDS 2002 reproduced with permission)

The concept of the file format is not discussed explicitly within the OAIS reference model. It occurs instead as a function of the transition from bit to digital object to data object, or as a feature of representation information. As such, it can leave several levels of recursion unarticulated.

Responsibility for the apparent terminological disconnect with OAIS in most of the literature on file formats lies with the need for OAIS to be fully comprehensive and its dating from the late 1990s - a case of definition more than criticism. Whether the ongoing revision of ISO14721 ought to be used as an opportunity to address this is explored later in this report. Three particularly stark examples illustrate the tendency. Firstly, the word *format* was used 115 times in the White Book version 3.0 of the OAIS Reference Model and mainly to refer to media format[3]. Secondly, OAIS does not assume that data conforms to a discrete file format specification. Thirdly, when it comes to discuss the related issue of *migration* (as in migration from one file format to another), OAIS uses the term *transformation* (the term *migration* is also deployed in OAIS to refer to media refreshing, replication, repackaging and other types of transformation, in addition to format transformation).

The inclusion by several sources of data streams in their definition of file format is possibly motivated by the generic nature of the OAIS model at these levels. This tendency should arguably be deprecated as inaccurate and confusing different levels of data and information encoding: it certainly does not belong in the file format literature. Whilst it is arguable that data streams may sometimes behave in analogous

[3] The OAIS glossary term from v.3.0 of the White Book (1998) defined ***format*** as: *The sequential organization of data in terms of its components*. This includes physical storage on media as well as encoding within a format specification (as in file format). This glossary definition was apparently removed during standardisation and does not appear in the 2002 Blue Book / ISO14721 although the same concept is present

ways to file formats, the need for clarity on how they are to be supported as information objects by representation information and preservation description information is not well served by this usage. McLellan (2007) notes the widespread confusion over textual encoding standards such as XML: this is particularly significant if the preservation community is to increase its influence in a web-based world. This issue is discussed briefly later in this report.

Direct consideration of information format by the OAIS Reference Model in the literature on file format selection is rare. Only Abrams (2007) and Christensen (2004) make much mention of OAIS and the latter only at a very high generic level: Abrams contains a full and authoritative statement of how OAIS builds up its information model from bitstreams on media right up to human interpretation through a series of potentially recursive layers in a context explicitly relevant to file formats.

1.5 A theoretical "minimal redundancy" paradigm

Information creators normally use technology at their disposal. Desktop applications often provide tools that are superfluous to a particular task. As an example, this report has been written using only a small subset of the capability of Microsoft Word. This makes inevitable a modest dislocation between the intellectual intent of a creator of a digital object and the code of the file format which contains it.

This dislocation has to be managed. Consider for a moment an impossible ideal where the file format specification and the intellectual object are perfectly aligned: the latter uses all the code of the former at least once and no code is redundant. It would be difficult to imagine that the preservation challenge of this scenario could be met by anything less than technological preservation or reverse emulation of the software supporting the format. In those practical respects the challenge is immense, but from the point of view of defining what it is, it could not be simpler: using several OAIS terms but not OAIS logic, preservation ought in principle to need no additional preservation description information (PDI), the representation information being entirely sufficient[4].

This ideal "minimal redundancy" paradigm will be revisited at several points in this report to clarify the issues then under discussion. The characteristics of a file format may be varying degrees of remoteness from the characteristics of the information it contains. Preservation activities ought to ensure that this remoteness does not jeopardise the human understanding of the information object.

[4] PDI is considered later. In all respects other than as a logical reference point, this paradigm is highly problematic, likely to tend towards a proliferation of niche formats and insurmountable resourcing consequences for repositories

2 Recommendations and conclusions

2.1 Conclusions for repository managers

2.1.1 It is not possible to recommend a definitive list of preservation formats for all data types in all repositories. However it is possible to establish a framework for the balancing of file format selection criteria which can be used to help repositories control and manage data deposited with them;

2.1.2 A review of the literature shows that there is consensus on five core and four wider file format selection criteria. Differences of application, detail and weighting can be resolved with reference to a repository's preservation strategy;

2.1.3 The core criteria influencing the choice of file format are: adoption, platform independence, disclosure, transparency and metadata support. Other considerations which should be borne in mind include: reusability / interoperability, robustness / complexity / viability, stability and IP / rights management;

2.1.4 Repositories should also factor in the extent of formats (the amount of media storage they require) and overall cost, although these are rarely articulated in the literature;

2.1.5 Beyond and potentially over-writing the criteria identified and cited above, repository managers should align the recognition and weighting of criteria with a clear preservation strategy that articulates the purpose of the repository and the needs of its designated community;

2.1.6 Depending on the nature of the relationships between information creators and repositories, repository managers may need to align their requirements with institutional IT practices and may wish to influence them accordingly. For example the adoption of proprietary desktop applications within an organization is likely to create a different risk profile for format management than the adoption of open source or non-proprietary software;

2.1.7 Risk management techniques should be deployed to ensure the continuing relevance of the strategy and the criteria to the collection being preserved. The strategy should be kept under regular review to ensure that it is still fit-for-purpose.

2.2 Recommendations for the wider preservation community

2.2.1 *For funders and policy makers:* Repositories need in the near future to have modelling tools and methods for the calculation of comparative costs of format choices across large collections. This is essential if cost is routinely to be factored into and balanced against other concerns;

2.2.2 *For researchers and developers:* Integrating the ability of formats to represent information content into scoring criteria seems some way off except for very simple digital objects;

2.2.3 *For researchers and developers:* The proposals of the archival community referenced in this report need to be translated and digested into the significant properties discussions in the wider digital preservation community and computer science. It is hoped this report has made a contribution to that process;

2.2.4 *For researchers and developers:* The development of non-proprietary file format specifications has had an effect on the data being presented to repositories. Scrutiny of emerging open formats, participation in their development and advocacy of their use is likely to reduce the risk profile of repositories;

2.2.5 *For those developing standards:* The concepts of provenance and authenticity in the archival community – represented in this report by discussion of InterPARES2 and InSPECT findings – seem to require some cognisance of the creator's intention in creating an intellectual object. This may need to be incorporated into OAIS, at the very least as a projection loop from the Producer to the Designated Community;

2.2.6 *For developers:* Some common approach to the description and discussion of preservation issues with web-based formats would be timely. There is a tendency in many quarters to focus on encoding languages rather than the higher level content representation issues and their technological dependencies. These are vital to resolving display and interaction issues in a preservation environment, particularly of complex, multi-object web applications;

2.2.7 *For researchers and standards developers:* The community should consider whether representation information and significant properties can / should be defined as exclusive categories (this has consequences for the development of metadata schemas and professional discourse). If not a pragmatic decision could be made to include as much as possible in representation information, based on the greater likelihood of automating its collection at present;

2.2.8 *For teachers and those disseminating research:* Significant investment of time is currently required to engage with the research literature in this area. This is in part a consequence of research projects operating at the cutting edge and having to define terms afresh, but other, more basic terminology is also used loosely within the community. For example: digital object, conceptual / information object, record. This can be problematic in itself, but makes exchange with other disciplines still more difficult;

2.2.9 *For researchers and those funding research:* OAIS concepts are helpful to much file format discourse, raising the question of whether key OAIS terminology could usefully be revised to make this clearer. The first downside of changes to OAIS terms to avoid conceptual clashes is that the reference model operates at a different level and this has its benefits. The second drawback is that, as in the delineation of significant properties / representation information and both of these / content information relative to preservation description information, sustaining a looser coupling is the only realistic option. It is recommended that at the very least, research proposals and papers map their information models carefully and explicitly against the OAIS information model to promote understanding and portability;

2.2.10 *For policy makers and leaders in the preservation community:* change in file formats remains a threat and drives up costs for long term access. The needs of the digital preservation community need to be represented to software vendors to reduce the risks and costs associated with the churn of file formats;

2.2.11 *For policy makers and leaders in the preservation community*: consideration should be given to a clearer statement on the benefits to long term access that accrue from the wider adoption of non-proprietary formats;

2.2.12 *For funders and those disseminating research:* It is recommended that the feasibility of a maintained digital preservation vocabulary is examined, with particular stress on keeping discourse mapped to OAIS. Most existing vocabularies have not kept in step with developments in this area. The Glossary to this report highlights many of the problem terms in italics.

2.3 Other conclusions from this study

2.3.1 The archival science discussion in this report further reveals the boundaries between preservation description information and information *content* itself being porous as they relate to dynamic and interactive content. As the community tackles these demanding environments, it will need new ways of expressing this complexity;

2.3.2 The findings of this report were arrived at independently but while writing a number of important papers have been presented that contribute to the discussions on significant properties and representation information (inter alia Dappert and Farquhar 2009, Giaretta et al 2009). Readers should be aware therefore that this is an active area of research.

3 Current file format recommendations

3.1 Methodology

Seven main sources which centre on the discussion of preservation file format selection criteria recur in the recent literature on file format selection, and are referenced comparatively here: Brown (2008a), Arms & Fleischhauer (2005), Rog and van Wijk (2008), McLellan (2007), Christensen (2004) and Huc *et al*. (2004) and Stanescu (2004). McLellan comprises a report based on 21 file format recommendations from either collecting repositories, research projects or leading libraries and archives[5]. Other sources, notably Abrams (2007) and Brown (2008b), discuss the issue of format selection very much from the point of view of representation registries. This section attempts to bring together the criteria proposed by these sources in a single place, discuss differences of emphasis and link to the wider preservation literature. Selected sources from McLellan have also been reviewed for additional detail using one of her InterPARES2 working drafts. A summary of the grounds of agreement is in the Annex, showing how criteria in these sources can be mapped against one another with very little adjustment.

3.2 Normalising the discussion and the extent of consensus

Several of the sources use a two-level hierarchy comprising criteria and sub-criteria – an example of this is given in section 3.5 in the discussion of Rog and van Wijk's scoring method. Sub-criteria are generally used to articulate detailed considerations within the issues represented by their "umbrella" criterion and to achieve aggregate scores taking this complexity into account – this too is considered later. Some use different terms to mean the same criterion, the most common example being *documentation* and *disclosure.* Occasionally, a source treats as a main criterion something that another considers to be a sub-criterion of something else or decomposes a criterion further than others see as necessary. The sources' rationales for their criteria have been examined to map their intention and the meaning as well as their main criterion terminology is shown in the Annex. The table in 3.5 shows how with the addition of a set of "candidate" main criteria and a small amount of merging and splitting of cells further down to accommodate differences of emphasis at the main criteria level, a consensus of five main criteria can be observed. They are:

> Adoption – the extent to which the format is in widespread use
>
> Platform independence- the extent to which the format is independent of specific support from hardware and software
>
> Disclosure – the extent to which the file format specification is in the public domain;
>
> Transparency – the readiness with which the file format can be inspected or interrogated to discover its identity and attributes, as against where it is obscured by compression, 'wrapper' data architectures or other techniques;
>
> Metadata support – the extent to which descriptive information is supported in extractable form within the format. This includes OAIS representation information and occasionally how far the file format supports the recording of

[5] McLellan included a significantly different earlier version of Brown (2008a): http://webarchive.nationalarchives.gov.uk/20060820092744/http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_file_formats.pdf and a Library of Congress webpage version of Arms and Fleischhauer http://www.digitalpreservation.gov/formats/sustain/sustain.shtml, but this element of double-counting is not significant

management processes it has been subject to (e.g. Microsoft Office document properties).

Even at this high level, there are tensions between these considerations: this is discussed generally in later in this section and as it relates to the specific tension between adoption and disclosure, and the slightly different issue of standards in section below. If the issue of standards is related to platform independence and central to *adoption vs*. *disclosure*, the remaining core criteria are technical issues to do with the structuring of the format itself. *Metadata support* concerns the ability of the file format to be self-describing. This can cover a range of descriptive ability, from representation information to management information. The latter is described by some library sources as support for authenticity (Christensen is an example noted in the Annex), something the archival community would find it hard to agree with. One aspect of *transparency* is most eloquently expounded by Arms and Fleischhauer:

> *digital formats in which the underlying information is represented simply and directly will be easier to migrate to new formats, more susceptible to digital archaeology, and allowing easier development of rendering software ..... Transparency is enhanced if textual content (including metadata embedded in files for non-text content) employs standard character encodings.... and stored in natural reading order ..... encryption is incompatible with transparency; compression inhibits transparency.....*

Simple and direct representation is also hampered by encapsulating formats inside a wrapper format, where they can be difficult to detect and manage.

3.2.1 Rearticulating the "Standards debate" in terms of adoption and disclosure criteria

McLellan found much vague recommendation of the use of standards and clarifies helpful distinctions between formats of non-proprietary origin and ones with open specifications (disclosure). She references OCLC and Cornell University agreeing with Brown[6] that the quality as well as the availability of documentation can be significant. Arms and Fleischhauer see the impact of patents, rather than proprietary origin as a potential independent factor: the point is well made as, all other things being equal, unencumbered use is the desired result.

It may also be helpful to view the openness / adoption tension as a preference for disclosed, *de facto* and *de jure* standards. The first two are blind to origin of a specification but concentrate instead on its documentation (openness or *disclosure*) and its (*de facto*) adoption. A *de jure* approach relies on the mandate and influence of the issuing organisation, be it a sectoral, national or international standardisation body or a leading practitioner organisation such as a national library or archive. Behind these mandates there is likely to be an interest in stable and managed development of file fomats. McLellan references several authorities (OCLC, MIT, Library of Congress, Cornell) agreeing with Brown[7] that wide adoption will produce market pressure for converters and other tools to keep files in such formats usable.

Many repositories' policies insist on open standard formats for AIPs and mandate them for SIPs, yet a significant number do not. The preservation research community has a natural interest in standards as providing a defined, disclosed and comparatively stable target for developers of software tools for our major processes of creation, migration and access. This interest is confirmed by the existence of two DPC Technology Watch reports examining the preservation potential of two ISO standards:

[6] The previous version of Brown 2008a, Op Cit
[7] Op cit

PDF and JPEG2000, see Fanning and Buckley (both 2008). McLellan's summary recommends the use of widely adopted formats if suitable open ones are unavailable.

3.3 Outside the core criteria

Beyond this core, there are five more criteria that are mentioned frequently but not universally. These are, unsurprisingly, articulated more diversely than the core.

> *Re-usability / interoperability* – the extent to which the format is interoperable with software, services and tools, enabling the content to be manipulated and reused for new purposes.
>
> *Robustness / complexity / viability* – Huc *et al*. consider a simple format inherently more preservable (a hint at the unachievable "minimum redundancy" paradigm?), but then include a separate criterion requiring a format to be capable of representing the full richness of content (the normal usage of *complexity*) and also consider the extent to which the format is resistant to corruption through internal error correction techniques / ability to recover itself from single points of failure. This is a criterion with contrary opinions. Others such as Brown (2008a) and Rog and van Wijk see complexity as a good thing if it includes error correction facilities
>
> *Stability* – The extent to which the development of the format follows a managed release cycle and provides backward compatibility. Amongst the sources cited, this is stressed only by Brown (2008a) and Rog and van Wijk (2007, but articulated as a subset of robustness) but is particularly significant with proprietary formats.
>
> *IP / Rights management* – The extent to which the format supports the management of intellectual property rights of either the preservation repository or third parties and, conversely, the extent to which it is encumbered by protection (e.g. by inhibiting copying and other reuse).

It is clear that some of the divergence at this level reflects different types of repository, particularly archives *vs*, digital libraries. Some of the more interesting divergences of opinion on these are discussed in the next section. A couple of the sources cited in the Annex also mention other criteria: Christensen's *simplicity* criterion includes simplicity of understanding, implementation and description, for example.

3.4 The 'absent' criteria: cost, extent and ability to represent full content

Three major considerations that might on inspection be expected to feature prominently as selection criteria were mostly absent from the literature reviewed:

> *Cost* requires enquiry and modelling activities beyond the scope of this report but closely linked to the recommendation on the role of preservation strategies: the cost and frequency of format migrations, their triggers and techniques for comparing the total cost of maintaining existing formats against these interventions;
>
> Managing metrics for *Extent* are clearly part of the bread-and-butter of many repository managers' current activities but it is surprising that it rarely appears in this context. Again there is a strong link with preservation strategies as well as the broader OAIS issues of producer and designated community interests: for example some large image archives are taking an interest in the emergence of the 'virtually lossless' JPEG2000 format (see Buckley [2008]), owing in no small part to the extent of uncompressed TIFF. Similarly a broadcast media archive is unlikely to operate without the use of compression formats such as MPEG;

The issue of *Content representation capability* was briefly discussed above in the consideration of complexity and simplicity in the 'wider' criteria. A detailed examination of this issue from an archival perspective is contained in this report.

These are criteria that will for many repository managers be more compelling than the 'core' ones encountered in the literature, yet judgements about them have important consequences for the operationalisation of those criteria. It is recommended that the preservation community undertakes cost modelling enquiry before consideration of these issues has a hope of being integrated into the scoring of other criteria in the future.

3.5 Grouping, weighting and hierarchy of criteria

In the literature-based review and conceptual analysis already cited, McLellan was categorising the criteria of institutional repositories and research initiatives. Most of these - 16 out of the 21 surveyed - are behind deposit requirements or recommendations of repositories. Rog and van Wijk are very honest in describing their perspective as informed by the progress of the e-Depot from e-journals and digitisation activity[8] towards a broader repository but make an interesting statement:

> *as the weighing of these criteria is connected to an institution's policy, the KB wonders whether agreement on the relative importance of the criteria can be reached at all ..... the examples in this paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense. (Rog and van Wijk 2008, 1)*

This report agrees strongly with this statement, which is its primary finding. The consequences affect both the grouping of criteria (as they can heighten or eliminate the significance of issues) or how the groups themselves are balanced against one another. This is explored next.

A number of the selection criteria appear in some sources at a slightly different level: this is annotated in the Annex. For example, Rog and van Wijk identify many of the same criteria as other sources, but broken into sub-criteria and with the addition of a scoring method by main criterion. For example, the following table is their previously cited decomposition of ***robustness***:

Robustness		
Format should be robust against single point of failure (2)		
	2	Not vulnerable
	1	Vulnerable
	0	Highly vulnerable
Support for file corruption detection (2)		
	2	Available
	0	Not available
File format stability (2)		
	2	Rare release of new versions
	1	Limited release of new versions
	0	Frequent release of new versions
Backward compatibility (2)		
	2	Large support

[8] They offer - p. 5 - an interesting example where a partner organisation prefers a format they score low on account of its dissemination capabilities

	1	Medium support
	0	No support
Forward compatibility (2)		
	2	Large support
	1	Medium support
	0	No support

In this example, it is the sub-criteria rather than the main criterion ***robustness*** that have a weighting score (2). The actual sub-criterion scores are applied on a scale of zero – 2 prior to the weighting factor being applied. Grouping criteria and their scoring against one another is an important issue. If aggregated scores of 'related' criteria are used to determine preservation decisions, aggregating methods can affect the arithmetic and the outcome. The upshot of this is that grouping is itself a weighting technique and it too needs to be considered in the light of preservation strategy. This rather undermines the potential of breaking criteria into sub-criteria and orienting scoring scales to resolve some of the contradictions outlined above.

The use of different preservation solutions and formats should not be an *ad hoc* decision looking only at the instances of the present format and assessing likely target formats against such criteria. The OAIS model demands the assessment of the information object delivered through the DIP according to the designated community's needs and few are in a position to do this without a mind to its technical and economic feasibility. Demands from various quarters for the preservation of content also requires governance of preservation processes such as migration. Typically, this will cascade down from a high-level policy through a clearly articulated preservation strategy.

4 Reconciling contrary criteria and scores

The tendency for preservation file format selection criteria to contradict one another has already been touched on in a previous section on the 'core' and 'wider' criteria. Particular preservation environments may make a number of criteria more or less significant. For example, the type and extent of collaboration with the producer and designated user communities may determine the scope and requirement in terms of the adoption of specific preservation actions by creators or the need or absence of demanding dissemination requirements. A few other sources were reviewed beyond those in the Annex, but the tendency to diverge is even more marked. This is apparent in Christensen, whose concerns are very specific to web archiving formats.

4.1 Divergent criteria

Some examples of divergent criteria are worth discussing:

adoption vs. documentation (*openness* / *disclosure*) Repositories such as some public archives receiving most of their content direct from standard desktop computing environments may receive a high proportion in proprietary Microsoft formats, such as Word. The influence on producers of preservation considerations may be limited, even if there is overlap between the producer and designated communities. Many sources express a decisive preference for open formats – some such as Huc *et al*. insist on it. It is encouraging that there are currently (2008-09) strong tendencies for proprietary formats to become more open and - perhaps - for open formats to become more adopted. Examples of the former are the PDF family of formats and Microsoft Office Open XML format.

simplicity vs. complexity This is a difficult area. Common-sense says that a simple format ought to be easier to preserve owing to a lower platform dependency, possibly likely to be more transparent. On the other hand, a complex format might allow the richer representation of a wider range of content or provide internal validity and integrity checks that would otherwise have to be carried out by external tools.

transparency This criterion is usually associated with the ease of accessing information held within the format, be it representation or content information. The examples usually given are compression, 'wrapper' data architectures and encryption / other deployment of digital signatures: access to the correct algorithm or contained objects are additional dependencies preservers could often do without. Some types of content, such as large static or moving images, are so extensive to make deployment of these techniques inevitable (see previously on the 'absent' criteria for further discussion of this).

usability / interoperability This is another criterion which is not universal and is most prevalent where information reuse is important. It may simply be a question of the generation of DIPs, including the production of redactions masking or removing sensitive information. There are a range of issues occurring far less frequently and relating to usability, often phrased as interoperability, ease of rendering, "manipulability" and such terms. These tend to occur where there are obvious links to the business model of the recommending institution, such as open archives. The link to repositories' preservation strategies, which obviously need to be linked to their broader business and service models is discussed below.

An example of such a criterion in action is the comparison of an image format with some textual support – such as PDF - with a more straightforwardly text format. PDF is favoured in some quarters as providing the 'look and feel' of

a straightforward document as a picture. Against that need to be set the need for particular software to search or extract text.

Digital rights management (DRM): for and against Scoring this criterion could work in contrary directions, according to issues that should be addressed clearly in the preservation strategy. A repository, particularly a small digital library, may have a particular business requirement for DRM capability. If it has an 'open archives' policy, it may specify its AIP format in SIP conditions, eliminating a migration step. Another repository could plausibly see the presence of DRM features in AIPs (or SIPs) - with the additional dependencies that DRM involves – as wholly detrimental.

degree and type of metadata support Although this has already been mentioned as a core criterion, at this level there is also divergence. This is in part owing to a distinction between those providing producer guidance and those stating internal repository criteria and variations of opinion on whether this is best achieved inside the file format and at what stage it is to be extracted. Other differences arise according to whether a repository uses a relational database or embedding techniques to manage representation information[9].

4.2 Role of preservation strategies

Two questions arise for the preservation community and repository managers from the foregoing discussion of file format selection criteria:

- How far do the divergences, different emphases and nuances within the five core and four wider criteria matter?; and
- How is a repository to make clear decisions on preserving information content based on such a bewildering number of considerations as proposed by these nine and other, more localised considerations?

Indeed, it is quite possible to devise new detailed criteria of relevance in a particular preservation scenario – there are plenty of signs in the literature of an ongoing proliferation of criteria already. This proliferation does not necessarily have to be a source of angst, though it does make a definitive list, grouping and scoring method very problematic. This report proposes that beyond and even within a few core criteria already identified by the community and cited above, the most important action is to align the recognition and weighting of criteria with a clear preservation strategy and keep them (and it) under review using risk management techniques. Evaluating and managing the risk to the preservation of content objects and the developing profile of the collection bearing in mind the stakeholders of the repository (including the OAIS designated community), its resources and mandate will involve regular review of how this is working in practice. This would amount to applying standard risk management techniques to the preservation challenge[10].

As long ago as 2005, Andres Stanescu proposed a similar risk-based approach to managing preservation risk. His INFORM methodology addresses wider concerns

---

[9] The attributes of format contained in representation information has, an extensive literature and one authoritative source on file formats devotes a startlingly high proportion of his space to discussing representation registries (Abrams). The principal other references are Brown (2008b) and PREMIS. This issue is treated as out of scope in this report, except here and in the discussion of preservation description information / significant properties below. Representation information needs to be adequate to support data management functions and allow the application of preservation strategy and criteria at the appropriate stage

[10] The DRAMBORA method for assessing repository activities more generally against the CRL/OCLC: Trustworthy Repositories Audit & Certification (TRAC) check-list is compatible with this approach. See http://www.repositoryaudit.eu/ and http://www.crl.edu/content.asp?l1=13&l2=58&l3=162

than file formats and issues handled by the other sources as file format risks (software and hardware dependencies, interoperability risks between the repository and producers / users and migration risks) are separated out from file formats. The method would seem to be partially refactored by the consensus on five format selection criteria noted above, but the sharing of knowledge on particular format risks and its higher-level aggregate scoring according to a repository's other risks is methodologically consistent. Dappert and Farquhar's (2008) organisational modelling technique within the PLANETS Project seems to be a revival of this (see below).

4.3 Revisiting some prominent digital preservation strategies

Familiar and broad types of preservation strategy, specifying the timing and type of preservation action – 'at' or 'prior to creation', 'at ingest' 'on-demand' or 'at obsolescence' migration, normalisation and various combinations of these - will impact significantly on the selection of formats, as will the objects comprising the collection, the relationships with the producer and designated user communities and a repository's resources. A preservation strategy needs to govern which approach is being mandated to assign governance to the scoring of criteria and implementation of preservation actions.

Hindsight permits the revisiting of some pioneering preservation strategies in the light of the foregoing discussion of format selection criteria. The institutions concerned might articulate their initiatives slightly differently if doing so today.

4.3.1 National Library of the Netherlands

Rog and van Wijk's account of the broadening of the National Library of the Netherlands' digital collection requiring a review of preservation strategy is instructive. A progression from digital publications preferring mainly PDF to scientific publications with associated proprietary office formats and on to archived websites has driven reconsideration of an approach based on normalising to PDF/A towards accepting a far wider range of formats.

4.3.2 Public Record Office of Victoria

At the Australian State level, the Public Record Office of Victoria has built its preservation activities (Victorian Electronic Record Strategy, or 'VERS') around the VERS Encapsulated Object (VEO) since the late 1990s. The VEO uses PDF renditions of record content and metadata wrappers in a distinct architecture. The strategic decision to deploy PDF was apparently taken at a very early stage based on accuracy of rendering of records with a traditional documentary character[11] and to merge AIP and DIP formats.

4.3.3 National Archives of Australia

The National Archives of Australia (NAA) announced its Xena normalisation strategy in 2002 as a groundbreaking one designed to preserve the "performance" of the record by encoding and wrapping in eXtensible Markup Language[12]. The Xena software has been freely available but few other repositories have followed NAA's lead in using it wholesale. The early press for Xena stressed the use of XML as the headline rather than the present focus on migration from proprietary format to the nearest OpenOffice.org equivalent[13]. The ease of use of Xena, the free availability of the software and most importantly the quality of renditions from proprietary office document formats then in use certainly seemed to promise much. The subsequent wider adoption of open source desktop applications such as OpenOffice.org and the

[11] http://www.prov.vic.gov.au/vers/standard/advice_13/. Similar simple digital records are considered by the InSPECT Project, see section 5.2

[12] The NAA 'performance' model is discussed further in section 5.1

[13] See Heslop, Davies and Wilson, (2002)

open formats such as produced by the Organisation for the Advancement of Structured Information Standards may have altered the requirements originally posited in the XENA project.

5 Preservation tools and infrastructure to support strategy implementation

The development and use of tools developed within the digital preservation community has a mostly separate literature from that of defining and implementing selection criteria. Brown (2008a) and particularly Abrams are the main exceptions to this in their articulation of a tight linkage to representation registries. There is a broader range of relevant tools already available or under development to assist in the management of preservation risk including format identification[14] and validation[15] software. It is probably correct that the risks of formats do not, strictly speaking, reduce with the development of tools. Rather, the means of mitigating them are more widely available.

5.1 File format and representation registries

The role of representation information in OAIS is to accompany data objects with sufficient technical description to enable their sufficiency as content objects. Across time this means tracking the dependencies of current data objects to enable them to be migrated at the appropriate point to more current or stable technology. Registries such as PRONOM[16] and the planned Unified Digital Formats Registry (UDFR) [17] aim to do this by the recording of technological dependencies of file formats and dissemination, including through alerting services. The Registry Repository of Representation Information developed with Digital Curation Centre and CASPAR Project funding aims to tackle the full range of representation information[18]. Holding this representation information centrally in registries, against an identifier for the precise file format means that repositories do not necessarily need to store all the representation information in direct association with objects in their collections, but could opt to link to it through the registry's format identifier. The metadata they can then turn their attention to managing is then the significant properties of the instances discussed in the next section.

5.2 PREMIS metadata dictionary

PREMIS[19] is the leading initiative setting out the metadata implications of OAIS. Comparing it with early preservation metadata schemas, particularly those for digitised image files, shows how much the landscape has changed. There is an explicit driver in PREMIS only to define the core metadata applicable to all repositories and information objects: it explicitly rules out preservation description information needed to preserve intellectual objects, whereas some of those early schemas strayed into that area, apparently unwittingly. This issue of the dividing line with preservation description information will be also discussed.

5.3 Current European research outputs in characterisation and preservation planning

The EU-funded PLANETS project[20] is developing the PLANETS Interoperability Framework, with the aim of maximizing automation and scale of characterisation, preservation planning and execution.

Characterisation is the activity of determining an adequate representation network for digital objects. PLANETS is developing automated methods of extracting characteristics and building characterisation registries on top of format registries such

[14] For example The National Archives' DROID: http://droid.sourceforge.net/wiki/index.php/Introduction
[15] Such as Harvard's jHove: http://hul.harvard.edu/jhove/
[16] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[17] See http://www.udfr.org/
[18] http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/
[19] See PREMIS working group (2008)
[20] See http://www.planets-project.eu/

as PRONOM, as explained by Brown (2008b). It is also developing PLATO[21], a web-based preservation planning tool which can be used to input the sort of file format selection criteria already discussed, assign different weightings and test the outcomes. Planning decisions can be audited by the saving of the scoring method and its assumptions. This is already available for use at the time of writing, although arrangements for its long term future have yet to be announced - the maintenance of such infrastructure would fill a real need.

At the more strategic level, Dappert and Farquhar (2008) have set out a method for modelling organisational objectives and representing them in machine-readable ways. The model represents the many levels up from bytestreams, right up through file formats to intellectual objects, collections and environmental entities such as policies. The aim is to line up top-down policy and strategic issues with bottom-up issues such as file format criteria: this has the potential to provide a methodology for resolving tension between different types of considerations and risks. Their introductory discussion shows significant agreement with the risk-based approach to file formats discussed earlier.

5.4 Significant properties

To this point, this report and its cited sources have mainly been concerned with the preservation of content information objects (the data, plus its representation information). OAIS acknowledges the necessity of preserving information usable and understandable to the designated community. This dimension broadens the perspective radically. Considering only the content information - the digital objects plus the representation information – is inadequate to achieving this: Preservation Description Information (PDI) is also needed. This has very significant implications for digital preservation practice because the originating format is only a carrier for a message. On the one hand, there may be characteristics of the originating format that may be essential to the interpretation and use of the DIP (through the AIP) by the designated community and prone to being overlooked if not carefully considered. On the other, it may include in its specification many unused facilities that are not required in any target migration format, as discussed above.

[21] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

A great deal of research has been conducted in several information science disciplines to parse the abstract, conceptual or intellectual object or 'work' from the present means of manifesting it - these are particularly relevant to digital preservation[22]. Except in the OAIS information model, this has only recently begun permeating to the preservation community's research and practice. This report now goes on to consider how records and archival community research may have a major contribution to make to wider digital preservation thinking about this issue.

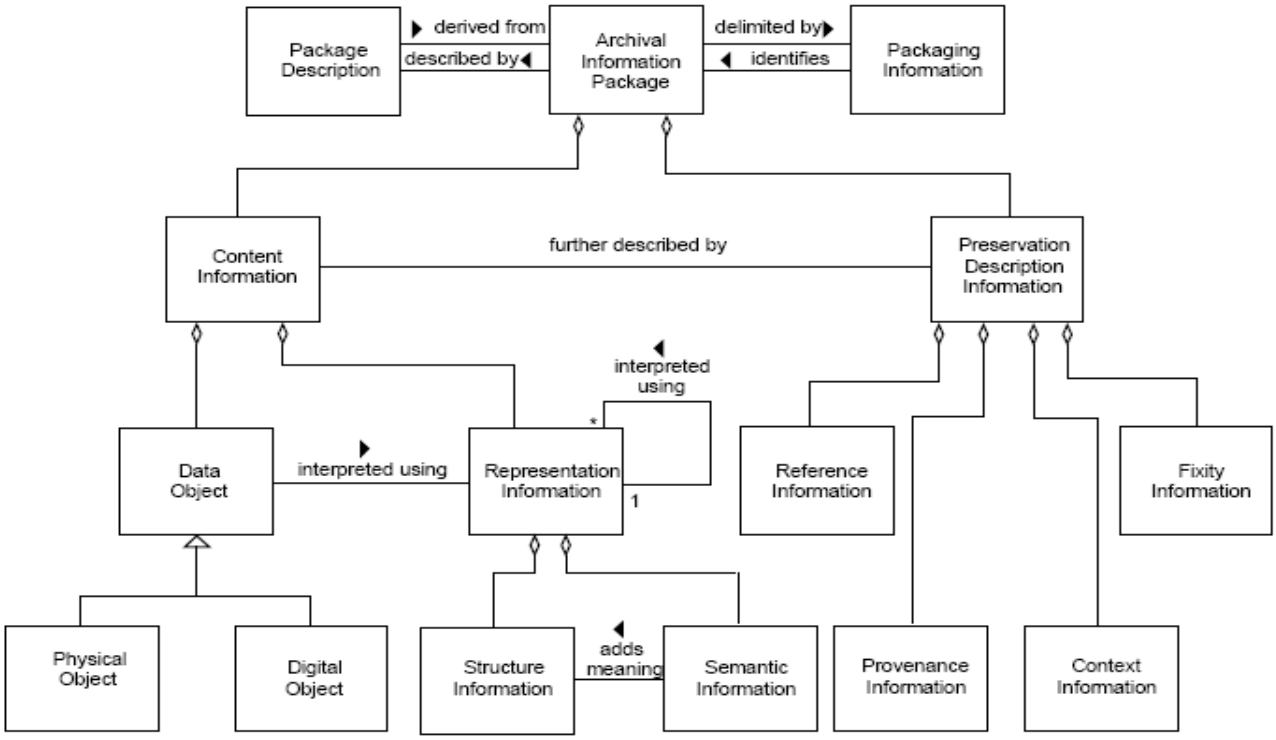

Figure 4-18: Archival Information Package (Detailed View)

Figure 2: OAIS context of preservation description information added to content information within the AIP (Figure4-18 within CCSDS 2002, reproduced with permission)[23]

[22] See IFLA Study group (1998)

[23] The perspective taken here by the OAIS model leads it to introduce an unfortunate change of terminology at a crucial point: what is *content information* on this view was called an *information object* on the lower level view extracted in section 1.4

6 Contributions from archival research and practice

The archival viewpoints considered in this section of the report are based mainly on archival specialisations of two issues identified by Clifford Lynch in 1999 and 2000. Lynch's own perspective is from the digital library community, but very aware of the equivalent concerns of other information science sub-disciplines, including archival science. Lynch (1999) proposed the 'Canonicalization' (here amended and anglicised to 'canonicisation') of the significant properties of digital objects to enable them to be measured. The InSPECT Project is a pragmatic approach to this for simple digital records proposed by a collaboration between the UK National Archives and the King's College London Centre for eResearch. In Lynch (2000), the issues of authenticity and trust in even dynamic, interactive and experiential environments are discussed. Such environments have been studied in a large body of case studies in the second phase of the InterPARES Project. Wilson (2007) provides a useful summary of significant properties initiatives in the intervening period, with a mainly but not exclusively archival perspective.

6.1 The 'performance' model, ISO 15489 and content vs. documentary form

The notion of a "performance" conveying the "essence" of a digital record has already been mentioned in the context of the preservation strategy of the National Archives of Australia and the issue has been taken up by several research projects. The generic NAA scenario, taken up by the *Inspect* project, is as follows:

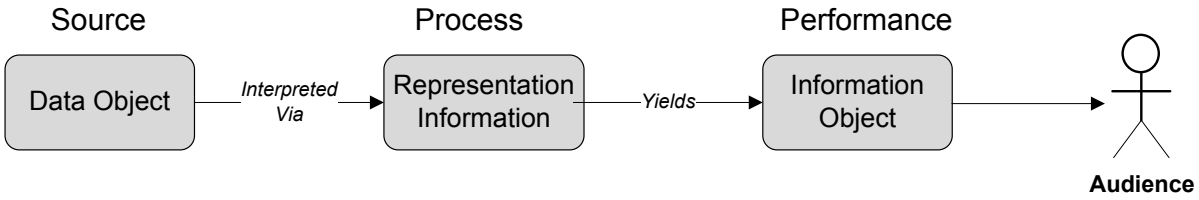

Figure 3: the application of the performance model to the recreation of an OAIS data object (Figure 2 in Knight 2008, reproduced with permission)

Re-examining the Australian Commonwealth example cited above, normalising to non-proprietary standards based equivalent of the submitted file - rendered in XML - in addition to maintaining the original bitstream and wrapping both in an XML metadata wrapper - assumed that the character of the information can be conveyed by the submitted format (supported by its representation information). Logically according to OAIS and a mixture of hindsight and advanced common sense this cannot be the case, although as with the VERS approach with apparently "traditional" records and even simple websites it may have appeared to be the case. For example, a single-item physical document such as a letter could be equated to a PDF containing the same appearance, textual content and structure in terms of paragraphing, salutation, heading, date, etc. without too much logical distortion: this issue is picked up in the following section on the *Inspect* project. The beauty of the "performance" model, though, is its clear extensibility into more complex technological environments, as will be observed later.

Examination of properties by their significance rather than their nature has two major effects:

- It causes the study of OAIS representation information and aspects of preservation description information to be evaluated in the light of their

precise contribution to the preservation of understandable, interpretable information; and

- It creates a tension between properties applying at the digital object against those relevant at the AIP level[24].

The logical interplay between a notion of 'performance' within *Inspect* and the focus on dynamic, interactive and experiential artistic data within the InterPARES2 project taking on board artistic views of their authenticity is fascinating, but consideration of it is best left until simpler scenarios have been considered.

6.2 InSPECT Project: Canonicising significant properties of simple digital records

The *Inspect* project aims to address the issue of significant properties by proposing canonical lists of criteria and measurement scales[25]. This is a very important issue: a measurable performance method is a necessary tool to defining acceptable loss. The scope of the project as set out in Wilson includes a useful survey of previous significant properties research in Europe, North America and Australia that is mostly superfluous to repeat here. The examples used to illustrate the *Inspect* method show a pragmatic address to the issues already faced by the partner institutions and many others like them: the illustrations are single file digital objects: vector graphic / raster images, emails, audio files and structured textual documents.

The other noticeable aspect of the method is to define significant projects in a way informed by a pragmatic archival viewpoint - the objects are at many points referred to as being "records". This seems driven mainly by the project partners' businesses and the observation of a wide definition of a record derived from the Australian archival tradition (i.e. from the Australian standard that underwent international standardisation eventually to become ISO15489 in 2001). The Project is explicitly aware of certain key terminological differences in this area and sets out to produce an analysis tree from the following top level structure, an approach derived from Rothenberg and Bikson (1999) as well as Lynch (1999)[26]:

- Content
- Context
- Rendering
- Structure
- Behaviour

Each record 'type' is described and assessed using an analysis template resembling a tree. the contribution of these properties to the information fulfilling its function as a record is evaluated. There will be a combination of technical and intellectual properties. The former often map directly to some of the file format selection criteria discussed earlier in this report and – ideally – recorded in representation information. Defining the latter is operating more at the cutting edge. Some may be acceptable ranges for representation information for the information concerned, such as resolution requirements in image files. The pragmatic use of groups of similar simple records as generic intellectual objects within InSPECT is the first step: the next might be to decompose these groupings into more specific intellectual objects from the point of view of the creating organisation's business: for textual documents, examples might include correspondence items, reports, minutes of meetings and so on.

[24] for clarity, digital object is used in the OAIS meaning here

[25] There were other JISC-funded studies of object-types not forming part of InSPECT: moving images, computer software, learning objects and vector images. See: http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx

[26] The method is an explicit implementation of canonicisation as proposed by Lynch (1999)

6.3 Archival science and the InterPARES2 Project: resolving the content and authenticity problems in complex dynamic, interactive and experiential environments

The InSPECT results from the structured textual and email documents show some of the same findings as the Authenticity Task force report of the first phase of the InterPARES Project[27], though they have been arrived at from a very different methodology. InterPARES1 examined "traditional" digital records from databases and document management systems and found that, with the "original record" disappearing as soon as it was saved from random access memory to other storage, it was only possible to preserve the ability to reproduce the record, not the record itself, providing a justification in archival science for migration of content. Its focus then shifted to methods of assuring the authenticity of those records through maintaining a stable documentary form.

In the case study environments of InterPARES2, the researchers had to contend with an even more fundamental problem: dynamic, interactive and experiential systems typically involve variable inputs from users and other programmes. This meant that the "fixed content" that in InterPARES1 could normally be assumed even when the file format of a bitstream had been changed was no longer present. This connects back to the discussions earlier in this report of the complexity criterion for file formats and the theoretical "minimal redundancy" paradigm.

The InterPARES2 case studies show investigation into very specific intellectual objects and ones that are technologically sophisticated. In 2002, Ken Thibodeau published a paper outlining the work of the first phase of the InterPARES project in articulating the implications of OAIS in the community concerned with preserving archival records (Thibodeau 2002). In 2006, well into the second phase of the research, he and Luciana Duranti articulated a conceptual extension of the definition of a record as understood by archival science and capable of accommodating records and other intellectual objects produced in the dynamic, interactive and experiential environments studied in the second phase (Duranti and Thibodeau 2006). They boil down to characteristics neither present nor explicit in "traditional" analogue or digital records: they have behavioural characteristics that vary according to some input from a user – including the end-user seeking only access - or another computer programme. Without intervention, they appeared to lack a fixed content (most also lacked a recognisable documentary form). In digital preservation rather than archival science terms they lacked most representation and any preservation description information. Worse still, from the point of view of the creators themselves they would probably be unable to fulfil their role as intellectual objects after any preservation intervention.

The particular innovations from Duranti and Thibodeau in response to this are the notion of "bounded variability" and the "prospective" as opposed to the retrospective record:

- A *Prospective record* comprises what InterPARES2 calls a set of digital components – including but not limited to many digital objects– and a set of instructions for the assembly of the intellectual object in the future. A traditional record – analogue or digital – is 'retrospective' in that it was normally definable as a by-product of a business activity. It had a fixed documentary form and stable content. These characteristics will not accommodate records from these environments where the digital components are capable of reassembly in other ways and the content has no fixed presentation or even data values.

[27] The InterPARES1 Authenticity Task Force was informed in part by diplomatic analysis: see McNeil, Gilliland-Swetland *et al*. (2000)

- *Bounded variability* is a measure of what Rothenberg, Inspect and others call "behavioural" characteristics. Variability in their presentation needs to be bounded by limits that would be recognised by and acceptable to the creator.

These are substantial departures for archival science. The necessity of detailed understanding of the creator's intent in creating intellectual objects is derived from both pragmatic and theoretical considerations. On the one hand the case study environments have their own demands that any preservation strategy would need to be measured against and current measures in place were found mostly to be inadequate. On the other, Duranti and Thibodeau were building on interdisciplinary comparison of concepts of authenticity conducted within the project[28]. This is based on extensive research data of 26 case studies in the arts, sciences and government and the case study data was collected in the former two categories irrespective of whether the entities were seen explicitly as records or not. In this context, and given the discussion of the 'performance' model used by NAA and *Inspect*, particular attention is drawn to the interactive electronic arts environment in the "Obsessed again" (hybrid musical work using a conventional bassoon and an interactive music programme) and "Waking dream" (dance performance involving robotics) case studies[29] .

Duranti (2008) has more recently gone further into the consequences for archival practice and articulated an augmented set of appraisal requirements. Appraisal of the feasibility of preservation now needs, she argues, to be an iterative process ideally beginning at or near system design or record creation stage in order to preserve authentic records from the sorts of environments studied. This will need to include the adequacy of file formats to represent content information. The *Policy framework and principles* also proposed by InterPARES2[30] detail the desirable relationship between creator and preserver to support this activity.

A record is a very specific content type with, in OAIS terms, a very stern provenance requirement. Some general points relevant to our main themes with application to other demanding intellectual objects in mind (such as digital artworks) can be drawn out:

- Firstly, these papers articulate for the archival community, its producers and designated communities a more detailed view of what OAIS states in very generic and high-level ways. OAIS states that how digital content is preserved and presented is not simply a technological problem and solutions need to be informed by preserving the intellectual object. Migrating a single content object from one format to another may be relatively simple compared to migrating digital objects formed of many discrete files. Interactions between digital components and with other agents need to be accurately preserved as well as the components' intrinsic characteristics: something that itself has an intellectual as well as a technical dimension;

- Secondly, the sort of interdisciplinary exchange facilitated within InterPARES2 between archivists, computer scientists, engineers, artists, public administrators and natural scientists throws up demanding quandaries. These have been articulated into archival theory but now need retranslation back into other communities' consciousness[31];

[28] See Roeder *et al.* (2008)
[29] See Fels & Danby (2007) and Amort (2007). Interestingly from an OAIS perspective, both also have traditional physical components. It will be very interesting to compare these case studies and the InterPARES2 project findings with the final products of the EU-funded CASPAR Project
[30] See Duranti, Suderman and Todd (2007)
[31] The initial scientific response to the interdisciplinary exchange with archivists inside the InterPARES2 project is offered by Laurialt *et al*. (2007). This is an essential companion to Roeder (2008) and Duranti and Thibodeau (2006) as it discusses key archival concepts from a scientific

- Thirdly, the inability of the current creators in the case studies to capture the content and representation information and the raising of the bar by the articulation of new authenticity requirements demonstrate how difficult this is to achieve with current capacity[32]. InterPARES2 was able to give detailed technical and conceptual attention to its case studies in a research environment. Operationalising the findings for a "real world" archive means automating this as far as possible. The PLANETS research mentioned earlier may be a step towards achieving this, but the preservation community is very much at the beginning of this process;

- Lastly, the prime importance for record authenticity of articulating the creator's viewpoint poses a serious logical problem for the OAIS reference model. The latter gives the producer responsibilities to provide or facilitate data objects, representation information and preservation description information, but the issue of authenticity is handled solely in terms of the requirements of the designated community – the users – insofar as it is handled at all[33]. At the very least, a loop projecting authenticity requirements from creators to inform the designated community is required in a revised OAIS to accommodate this.

InterPARES2 proposes neither a specific canonicisation of significant properties, nor a measurement method. Its findings come perilously close to recommending the preservation of entire systems and there must be doubts about the scalability of this approach[34]. It is interesting and heartening, though, to see that this very theoretical endeavour is producing some of the same findings as the more practical InSPECT approach[35]. Elsewhere in the InterPARES2 Project, it has issued the guidelines on file formats already referenced (McLellan) and guidelines for records creators but these do not work through the implications of these broader findings.

6.3.1 Broader application of archival viewpoints within digital preservation community

It has already been noted that the InterPARES2 findings need to be re-translated back into the discourse of other, non-archival communities for validation and use. The extended discussion which preceded is an attempt to do this in summary form for the wider digital preservation community. A perception may arise from the scope of InterPARES1 and the archival terminology and diplomatic methodologies within InterPARES2 that the latter's findings are not of interest outside archival science. There ought to be enough research data of interdisciplinary significance within InterPARES2 to counter such an objection. First and foremost, the scientific and artistic case study environments of InterPARES2 were interrogated as to their understanding of their "digital entities" (intellectual objects in this context) without an

---

perspective. The authors go on to make a series of policy recommendations to improve interdisciplinary collaboration and provision in Canada that have wider applicability.

[32] In one case study, the creators' business needs for extreme precision in engineering documentation across many decades had led to concerted efforts to describe the intellectual object, but still unsuccessfully. See Hawkins *et al*. (2007)

[33] In OAIS, PDI is composed of reference, context, fixity and provenance. These terms exist with other meanings in the archival community. There is a particular difference in the meaning of provenance

[34] Admittedly, preservation of the intellectual objects in many of the case environments according to the approach proposed by Duranti and Thibodeau (2006) could eliminate the need for the capture of individual "transactional" records in favour of preserving the means of reconstruction. InterPARES2 also took it as read that it should not generally take the expedient of static snapshots of dynamic data, although recognising that in a practical records management / preservation environment, this could be the best option

[35] It would be a mistake to see the conceptual foundations of InterPARES as identical to those of *InSPECT*. The prevailing viewpoint within the former is that of contemporary archival diplomatics, whereas *Inspect* is framed more in terms of ISO15489 and the Australian continuum viewpoint

insistence on their qualifying as records. Further, the importance of the interdisciplinary analysis of concepts of authenticity in the arts, sciences and government can hardly be overemphasized (Roeder *et al*. 2008). The extra-archival domains show equivalent concerns that require carful translation, but formal, philological and provenantial concerns are present there too. As an example, reuse of scientific data without regard for its lineage and the methodology used to collect it would not be acceptable in a scientific environment: this is not far removed from what an archivist would call provenance, business function, objectives and procedures.

The point can be generalised across other digital archival research as argued by Wilson from within the InSPECT Project. There, the methodological foundations are more closely related to the Australian records continuum where "recordness" is more a question of the careful attribution of context than something to be determined by inherent characteristics.

6.3.2 Extrapolating consequences of archival viewpoints into preservation file formats for distributed and web computing

A number of format-related preservation issues present themselves in these environments. The past five years have seen a remarkable increase in the use of the world wide web as a platform for far more than simply the provision of information and 'conventional' transactions. The web is in fact becoming both the development and the presentation platforms of choice for most and many activities cannot be carried out otherwise. An example within the InterPARES2 case study data is the "Cyber-cartographic Atlas of Antarctica"[36]: a complex case of a distributed geographic information system maintained by collaborators around the globe, none of them hosting a complete instance.

These tendencies present both opportunities and challenges for digital preservation. At the positive end of the scale, developing for web presentation at least implies observance of browser compatibility that ought to make accurate future presentation easier. Robust web protocols ought to provide some predictability of presentation if applied in both the creating and presentation environments, but the tolerance of many mass-market browsers for uncompliant code act in the opposite direction[37], as does the widespread use of short-lived browser plug-ins for multimedia content. This shifts the issues usually associated with file formats into a different space. It also introduces the tantalising prospect of the web acting as an emulation or even a virtualisation environment.

Again, stressing the positive first, changing business models in the ICT industry have seen the emergence of mass-market virtual applications associated with high-volume storage services. *Google apps*® is the most prominent example. These are typically far less complex than traditional proprietary software, with correspondingly simpler format specifications. This offers us a useful model for some of our own dissemination activity, where the format of such applications is of less relevance than their interoperability with many other delivery and dissemination formats. Users are typically able to render 'on the fly' to the format of their choice. Of more concern is the lack, so far, of any apparent intention to include preservation amongst the services offered.

Complex, distributed systems may well be best preserved *in situ,* not least because the maintenance of the constituent parts - including file formats and crucially the combined presentations and interactions between them - may have been specialised in

[36] Lauriault & Hackett (2007)

[37] The first attempt by the UK National Archives to archive the Number 10 Downing Street website in 2001 had to grapple with the obsolescence of a browser plug-in within 2 years of its implementation to run a virtual reality video file

the first place for good reason. This requires a different model from traditional custodial preservation institutions and possibly a different understanding of the lifecycle of digital materials. In *Managing the crowd*, Bailey suggests that new models and methods are also needed in the records management sphere to manage Web 2.0 content (Bailey 2008)[38]. Within the digital preservation community, there are already new models emerging such as federated and virtual repositories. If preservation *in situ* is preferable whilst there is an ongoing business sustaining the content, there may come a point where that is no longer the case and the only response to that may remain a more traditional custodial one[39].

The issue most closely related to preservation file formats, though, is that there may well be no definitive rendering of the content in the Web 2.0 environment – it is arguable that this is part of the definition of Web 2.0. InterPARES2 found instances in its case studies where there was insufficient metadata to establish the bounds of variability to consider digital objects as authentic records and made recommendations on how this might be captured, but arguably did not countenance an environment where no definitive rendering had ever existed.

There is also a tendency to overlook technological obsolescence issues provided one level of standardisation - usually textual encoding - is observed. Referring back to the generic OAIS model of information encoding layers, it is of limited assistance to preservation if our XML encoded format has no surviving or accessible specification. The principal identification method for a format on the web is its extension (HTML, XML, XHTML, etc.) as a file format and prior to the emergence of representation, but if the definition given at the beginning of this report is accepted this is incorrect. In some respects, they may resemble file formats but this looseness of definition obscures the need for attention to several layers of encoding, each of which may have obsolescence issues. This usage is to be deprecated. In addition, the advent of the semantic web means that a greater richness of the intellectual object and interactions between objects go far beyond static browser presentation issues. This requires an extension to McLellan's plea to keep document type definitions of XML objects to preserving the schemas, stylesheets and RDF bindings. The latter may encode business logic as well as behaviour: in OAIS terms these seem to straddle representation information and PDI. Their frequent composition from generic 'granules' referenced from collaborative registries also means that the preservation of their documentation is part of the same challenge. Perhaps the upshot of this is that the ability of cognate presentation with some metadata establishing previous usage are the absolute minimum. InterPARES2 requires the ability to reassemble the creator's intention within set bounds for this to be considered authentic, but concepts such as data lineage in the scientific community make similar demands.

[38] Bailey does not attempt to consider the preservation implications of his observations on records management

[39] Simple preservation services applied to individual files may well not lead to the cognate presentation of a digital object where the interaction between them or with the user needs to be choreographed

7 Defining significant properties: challenge of parsing representation and preservation description information

OAIS treats the categories representation and preservation description information as logically separate. As the foregoing demonstrates, they are anything but in the preservation community's literature and even some of the terminology we use to describe our current research activities. To be fair, cutting-edge research activity is re-examining these interfaces but one result is each project has had to define its own terms to carry out its research and another is that very significant engagement with research outputs is required to get a safe grip on what it is really being investigated. How far does this matter?[40]

Certainly, it is a higher priority that there is sufficient relevant information available to inform practical preservation decisions. A logical separation may only seem to be required in theory. Additionally, there are certain types of information that can seem to straddle the boundary or move across it according to the scenario. For example, if an attribute is generic to the current digital object type in an AIP, it is representation information and potentially a significant property of a technical object. If it is specific to the instance it is preservation description information and may also be a significant property of the intellectual object. A repository without detailed knowledge of the format or how it has been used by the information creator may not know which.

There are at least two compelling reasons why it is highly desirable for the community to clarify a theoretical distinction and agree a pragmatic working distinction. The obvious one is for the sake of our research activity. The second is because groupings of attributes will inevitably flow into metadata semantics and need ideally to be structured in ways that avoiding redundancy and clashes.

For the time being, it may be that representation information should take priority in the grey areas. It is at a more advanced stage of development and support by preservation infrastructure such as data dictionaries and registries and these offer more immediate hope of automation than those intellectual attributes we as a community have only recently turned attention to.

[40] At a significant properties workshop run jointly by the DPC and British Library in 2008, Kevin Ashley of University of London Computing Centre asked whether the community should be taking greater care in defining the relationship between significant properties and representation information. An important iPRES 2009 paper (Giaretta *et al*.) has been presented since the writing of this report surveying the usage of the term "Significant properties*"* and proposing instead a number of terminological innovations for the ongoing OAIS revision. The direct outcome is the proposed substitution of the term *Transformation information property for "Significant property",* but the authors also consider some of the same issues of provenance and authenticity present in this report

8 Conclusions and recommendations

The conclusions of this report and the implications for repository managers and the digital preservation community more generally have already been presented in the form of the Recommendations and Conclusions in section 2 of this report. There is little point in repeating these conclusions again, but it is worth re-iterating why the report has been structured in this way.

File format management is a topic of wide relevance. It impacts on the managers of trusted repositories most directly and through their recommendations and policies it impinges on anyone creating digital objects of lasting value. Consequently it is important that advice is lucid and timely. Yet the theoretical and practical foundations of format selection are not simple, nor are the archival and information paradigms from which such recommendations emerge. As we have seen, this topic has progressed rapidly in the last decade. This research has improved considerably our understanding of effective format management strategies – even if the proliferation of initiatives and tools seems at first to render it less accessible. Careful consideration of the issues has the potential to inform archival sciences in far reaching ways.

The primary finding of this report is to support the conclusion drawn by Rog and van Wijk in 2008 at the Koninklijke Bibliotheek (the National Library of the Netherlands-KB)

> *as the weighing of these criteria is connected to an institution's policy, the KB wonders whether agreement on the relative importance of the criteria can be reached at all ..... (Rog and van Wijk 2008, 1)*

Consequently, beyond and even within a few core criteria already identified by the community and cited above, the most important action is to align the recognition and weighting of criteria with a clear preservation strategy and keep them (and it) under review using risk management techniques

The thoughtful reader may now wish to return to the Recommendations for Action that follow the introduction to explore the implications in practice.

9 References

Abrams, S. (2007) *File formats* Digital Curation Manual instalment. Accessed from: http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats/file-formats.pdf, 15th May 2009

Amort, J.S. (2007) *Obsessed again* InterPARES2 arts case study. Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_cs15_final_report.pdf 15th May 2009

Arms, C. & Fleischhauer, C. (2005) *Digital formats: factors for sustainability, functionality and quality* IS&T Conference paper Accessed from: http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf, 15th May 2009

Bailey, S (2008) *Managing the crowd: rethinking records management for the web 2.0 world.* Facet Publishing

Becker, C. (2008) *The Planets Preservation Planning workflow and the planning tool Plato*, accessed from: http://www.dpconline.org/docs/events/080729becker.pdf, 15th May 2009

Brown, A. (2008a) *Selecting file formats for long-term preservation* The National Archives (UK) Digital preservation guidance note 1. Accessed from: http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf, 15th May 2009

Brown, A. (2008b) *Representation information registries* Accessed from: http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf, 15th May 2009

Buckley, R. (2008) *JPEG 2000 - a practical digital preservation standard?* Digital Preservation coalition technology watch report 08-01. Accessed from: http://www.dpconline.org/docs/reports/dpctw08-01.pdf, 15th May 2009

Caplan, P. (2009) *Understanding PREMIS* Library of Congress network development and MARC standards online Accessed from: http://www.loc.gov/standards/premis/understanding-premis.pdf, 15th May 2009

CCSDS (2002). *Reference Model for an Open Archival Information System* (OAIS). CCSDS 650.0-B-1 Blue Book Issue 1. Accessed from: http://public.ccsds.org/publications/archive/650x0b1.pdf, 15th May 2009

Christensen, S. (2004) *Archival Data Format Requirements.* Netarchivet, DK Accessed from: http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf, 15th May 2009

Clausen, L. (2004) *Handling file formats* Accessed from: http://web.archive.org/web/20051112022731/netarchive.dk/website/publications/FileFormats-2004.pdf, 15th May 2009

Dappert, A. & Farquhar, A. (2008) *Modeling organisational preservation goals to guide digital preservation* PLANETS; iPRES 2008 proceedings pp.5-12. Accessed from http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf, 15th May 2009

Dappert, A. & Farquhar, A. (2009) 'Significance is in the Eye of the Stakeholder' in M Agosti, J Borbinho, S Kapidakis, C Paptheodorou and G Tsakonas (eds) *Research and Advanced technology for Digital Libraries: 13th European Conference ECDL 2009 Corfu, Greece September 27-October 2 2009. Proceedings*, Springer, Berlin and Heidelberg, 297-308 online at

http://www.springerlink.com/content/e86474142788846t3, last accessed 27th November 2009

Duranti, L. (2008) *The Appraisal of Digital Records: Assessing More than Value* Conference of the UK Society of Archivists, York 2008 online at: http://www.archives.org.uk/professionalissues/conference2008presentations.html

Duranti, L Suderman, J. & Todd, M., (2007) *InterPARES2 Policy framework and principles* Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2(pub)policy_framework_document.pdf , 15th May 2009

Duranti, L. & Thibodeau, K. (2006) *The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES* Archival Science 6, 1: 13-68 (Online: http://dx.doi.org/10.1007/s10502-006-9021-7) Accessed 15th May 2009

Fanning, B. (2008) *Preserving the data explosion: Using PDF* Digital Preservation coalition technology watch report 08-02 accessed from: http://www.dpconline.org/docs/reports/dpctw08-02.pdf, 15th May 2009

Fels, S. & Dandy, S. (2007) *Waking dream* InterPARES2 arts case study. Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_cs13_final_report.pdf; 15th May 2009

Giaretta, D, ,Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G. & Sawyer, D. (2009) 'Significant Properties, Authenticity, Provenance, Representation Information and OAIS' in *iPRES 2009, The Sixth International Conference on the Preservation of Digital Objects: Proceedings*, California Digital Library, 67-73

Hawkins *et al*. (2007) *Preservation and Authentication of Electronic Engineering and Manufacturing Records* InterPARES2 case study. Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_cs19_final_report.pdf, 15th May 2009

Hedstrom, M. & Lee, C. (2002) *Significant properties of digital objects: definitions, applications, implications* DLM Forum proceedings, Barcelona 2002 pp. 218-237. Accessed from: http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf, 15th May 2009

Heslop, H., Davis, S. & Wilson, A. (2002), *An approach to the preservation of digital records*, Accessed from http://web.archive.org/web/20031217152126/http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf, 15th May 2009

Huc, C. *et al* (2004) *Criteria for evaluating data formats in terms of their suitability for ensuring long term information preservation* v.5 Groupe Pérennisation des Informations Numériques (PIN) Accessed from: http://www.ssd.rl.ac.uk/ccsdsp2/mon04/long_term_preservation_criteria.doc, 15th May 2009

IFLA Study group (1998) *Functional requirements for bibliographic records: Final report.* Accessed from: http://archive.ifla.org/VII/s13/frbr/frbr.htm, 15th May 2009

InterPARES2 Project (n.d.) *Terminology Database,* http://www.interpares.org/ip2/ip2_terminology_db.cfm, 17th November 2009

InterPARES2 Project (2006a) *Preserving digital records: guidelines for organizations*. Accessed from:

http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)preserver_guidelines_booklet.pdf, 15th May 2009

InterPARES2 Project (2006b) *Making and maintaining digital materials: guidelines for individuals* Accessed from: http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)creator_guidelines_booklet.pdf, 15th May 2009

Knight, G. (2008) *Framework for the definition of significant properties*, Inspect project Accessed from: http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf, 15th May 2009

Lauriault, T. & Hackett, Y. (2007) *Cyber-cartographic atlas of Antarctica* InterPARES2 case study. Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_cs06_final_report.pdf 15th May 2009

Lauriault, T., Craig, B., Fraser Taylor, D. & Pulsifer, P. (2007) 'Today's data are part of tomorrow's research: Archival issues in the sciences' in *Archivaria 64* 122-179

Lavoie, B. (2004) *The OAIS Reference Model: Introductory Guide* Digital Preservation coalition technology watch report 04-01. Accessed from: http://www.dpconline.org/docs/lavoie_OAIS.pdf, 15th May 2009

Lynch, C. (2000) *Authenticity and integrity in the digital environment: An exploratory analysis of the central role of trust* Accessed from: http://www.clir.org/pubs/reports/pub92/lynch.html, 15th May 2009

Lynch, C. (1999) *Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information* D-Lib Vol 5, No 9 Accessed from: http://www.dlib.org/dlib/september99/09lynch.html, 15th May 2009

MacNeil , Gilliland-Swetland *et al*. (2000): Authenticity Task Force Report, InterPARES Project. Accessed from: http://www.interpares.org/display_file.cfm?doc=ip1_atf_report.pdf, 15th May 2009

McLellan, E .(2007) *Selecting file formats for long-term preservation* (InterPARES2 project general study report) Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf, 15th May 2009

PLANETS project website: http://www.planets-project.eu Accessed 15th May 2009

PREMIS project *PREMIS data dictionary* (2008). Accessed from: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf 17th November 2009

Roeder, *et al.* (2008), *Authenticity, Reliability and Accuracy of Digital Records in the Artistic, Scientific and Governmental Sectors: Domain 2 Task Force Report*, Accessed from: http://www.interpares.org/display_file.cfm?doc=ip2_book_part_3_domain2_task_force.pdf . 15th May 2009

Rog, J. & van Wijk, C. (2008) *Evaluating file formats for long-term preservation.* Accessed from: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf, 15th May 2009

Rothenberg, J. & Bikson, T. (1999), *Carrying authentic, understandable and usable digital records through time* Rand-Europe report. Accessed from: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf, 15th May 2009

Stanescu, A. (2004) *Assessing the durability of formats in a digital preservation environment: the INFORM methodology. D-Lib Volume 10 N^{o} 11* Accessed from http://www.dlib.org/dlib/november04/stanescu/11stanescu.html, 15th May 2009

Thibodeau, K. *Overview of technological approaches to digital preservation and challenges in coming years* CLIR Conference paper 2002. Accessed from: http://www.clir.org/pubs/reports/pub107/thibodeau.html, 15th May 2009

Wilson, A. (2007) *Significant properties report* Inspect project. Accessed from: http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf, 15th May 2009

10 Table of core and wider criteria

	C o r e c r i t e r i a							*W i d e r c r i t e r i a*				
	Adoption	***Platform independence***		***Disclosure***		***Transparency***	***Metadata support***	***IP / DRM***	***Stability /backward compatibility***	***Robustness /Complexity / Viability***		***Re-usability***
(*)Brown TNA, UK (2008a)	*Ubiqity*	*Support*	*Interoperability*	*Disclosure*	*Documentation quality*	*Ease of identification and validation*	*Metadata support*	*IPR*	*Stability / backward compatibility*	*Complexity*	*Viability*	*Re-usability*
(*)Arms & Fleischhauer LoC, USA (2005)	*Adoption*	*External dependencies*		*Disclosure*	*Impact of patents*	*Transparency, incl.human readability;lack of encryption; natural reading order of textual files' content; standardisation of source code*	*Self documentation*	*-*	*-*	*-*		*-*
Rog & van Wijk KB, NL (2008)	*Adoption*	*Dependencies*		*Openness*		*Complexity*	*Self-documentation*	*Technical protection mechanism*	*Robustness*			
McLellan InterPARES2, CAN (2007)	*Widespread use*	*Platform independence*		*Non-proprietary origin*	*Availability of documentation*	*Compression*	*-*	*-*	*-*	*-*		*-*
Christensen Nerarchivet, DK (2004)	*-*	*Dependencies*		*-*		*-*	*Metadata support* / *Support for authenticity information*	*-*	*-*	*Robustness*		
Huc et al PIN group .v.5 FR (2004)	*-*	*-*		*Public standardisation*		*Inspectability*	*Extractability of metadata*	*-*	*-*	*Simplicity*		*Manipulability*
****Stanescu OCLC (2004)***	*Adoption*	*-*		*Disclosure*	*Documentation quality*	*-*	*Metadata support*	*DRM, signature, encryption facilities*	*Stability / backward compatibility*	*-*		*(as regards metadata interoperabilty)*

Notes:

i. With the minimal amount of merging and splitting of cells shown, 5 core and 4 other criteria can be identified from the listed sources. Minor differences of emphasis remain. Candidate titles for the ten criteria encompassing the meanings imparted across the sources are in the top row

ii. The starred entries are included in McLellan's survey but are either research-based and / or intended to have a wider application. In general, overlap between McLellan and the 20 institutional repositories whose documentation she summarises has been avoided. There is a small amount of double-counting remaining: Arms and Fleischhauer has a close relationship with the surveyed Library of Congress material and a substantially different previous version (2003) of Brown's UK viewpoint.

iii. Sources – many quoted by McLellan – focusing on digitisation formats or showing substantial impact from a distinct preservation strategy and curtailed consideration of a broad range of criteria have not been included here, e.g. DAVID Project, National Archives of Australia, Victorian e-Records Strategy

11 Glossary[41]

Term	Domain / reference	Definition	Comments
Adoption	Criterion	The extent to which a format is in use	
AIP (Archival information package)	OAIS[42]	An information package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved with an OAIS	
Authenticity	*Generic*	*Quality of being what is purported*	
Authenticity	*Archival science: InterPARES2*	*The trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that is free from tampering or corruption*	
Authenticity	*Records management: ISO15489*	*An authentic record is one that can be proven:* *a. to be what it purports to be* *b. to have been create or sent by the person purported to have sent or created it; and* *c. to have been created or sent at the time purported.*	
Behaviour		Characteristic of a file format involving either interaction with a user or another file	
Bounded variability	*InterPARES2; Duranti & Thibodeau*	*The limits to which interaction between users and components of digital records can be permitted to affect the behaviour of their composite Digital Objects*	
Bitstream	Wikipedia	A time series of bits	
Canonicisation	Lynch (1999)	"Canonicalization", (sic). Technique of defining irreducible rules, in this context generic scales for the measurement of Significant Properties	
Characterisation		Activity of determining adequate Representation Information (the OAIS Representation Network) for Digital Objects and collections of digital objects	
Complexity	Criterion	Extent of richness shown by a format in information representation	
Component	Generic	Digital Object	
(Component	*InterPARES2*	*InterPARES2 includes other 'digital components', e.g. Representation Information and PDI in its view of the Prospective Record)*	
Content information	*OAIS*	*The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information*	
Content object	*OAIS*	*The Data Object that together with associated Representation Information, is the original target of preservation*	
Creator	*Archival science*	*The legal or natural person whose business activities are recorded by a record*	*specialisation of Producer in OAIS*
Data	OAIS	A reinterpretable representation of information in a formalised manner suitable for communication, interpretation or processing	
Data Object	OAIS	Either a physical object or a data object	

[41] Italicised entries denote terms that are domain sensitive or problematic (novel, used loosely or the subject of disagreement)

Data stream	Wikipedia	A sequence of digitally encoded coherent signals (packets of data or datapackets) used to transmit or receive information that is in transmission	
Designated community	OAIS	An identified group of potential customers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities	
Digital migration	OAIS	The transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes: – a focus on the preservation of the full information content; – a perspective that the new archival implementation of the information is a replacement for the old; and – an understanding that full control and responsibility over all aspects of the transfer resides with the OAIS	
Digital object	*OAIS*	*An object composed of a set of bit sequences*	*See more specific definition of a File Format*
DIP (Dissemination information package)	OAIS	The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS	
Disclosure		File format selection criterion: extent to which a format specification is in the public domain	
File Format	*Abrams (2007)*	*a class of digitally-encoded assets defined by a set of semantic, syntactic and serialisation encoding rules for converting from abstract information to tangible byte streams*	
Global Digital Format Registry (GDFR)	Name	Collaborative project based at Harvard University with Mellon Foundation funding to build a global Representation Information registry	
Information	OAIS	Any type of knowledge that can be exchanged. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees celsius (the Representation Information)	
Information object	*OAIS*	*A Data Object together with its Representation Information*	
Inspect		Collaborative project between UK National Archives and King's College London Centre for eResearch to define Significant Properties of digital records	
Intellectual object			
Intellectual Property Rights management	Criterion	Extent to which – for good and ill – a format supports IPR protection and management	
InterPARES	Name	International archival research project based at the University of British Columbia, Vancouver. The first phase (1998-2001) examined the preservation of authentic digital records from document management and database systems, the second (2001-06) from dynamic, interactive and experiential systems in the arts, sciences and government. A third phase transferring knowledge into practice has begun and runs to 2011	
Manifestation		*Current OAIS content object within an AIP*	
Metadata	OAIS	Data about other data	
Metadata support	Criterion	Ability of format to be self-describing as regards Representation Information or management history	
*Migration**	*(OAIS*		*(See Digital Migration)*
MIME-type		Pragmatic identification scheme for file formats on the world wide web	
Open standard		Standard with a specification that is in the public domain	See Disclosure
PLANETS	Name	EU-funded research project	

Platform independence	Criterion	extent to which the format is supported by many different hardware and software platforms	
PLATO	Name	Product of PLANETS project	
PREMIS	Name	Collaborative project based at OCLC, aiming to work through the Representation Information implications of OAIS primarily through the production and maintenance of an authoritative data dictionary	
Preservation description information (PDI)	OAIS	The information which is necessary for adequate preservation of the Content Information and which can be categorised as Provenance, Reference, Fixity and Context information	
Producer *	OAIS	The role played by those persons, or client systems, who provide the information to be preserved	
PRONOM	Name	Representation Registry and associated services maintained by The National Archives, UK	
Prospective Record	*InterPARES2: Duranti & Thibodeau*	*Means of guiding future reconstruction of authentic records by the assembly of Digital Objects according to a set of instructions, including Bounded Variability. Developed as an extension to archival thinking by Duranti and Thibodeau after observing dynamic, interactive and experiential environments within InterPARES2 case studies and in contrast to more traditional Retrospective Records*	
Provenance Information	*OAIS**	*The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration*	*Archival use of the term is variant*
Record	Records management: ISO15489	Information created, received and maintained as evidence and information by an organisation or person in pursuance of legal obligations or in the transaction of business	
Representation information	OAIS	The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that determines how a sequence of bits (i.e. a Data Object) is mapped into a symbol	
Representation Network	OAIS	The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the long term	
Representation Registry		Repository of Representation Information typically linked to file formats to facilitate the tracking of technological obsolescence over time and assist preservation planning	
Retrospective record	*InterPARES2: Duranti & Thibodeau*	*Term coined by Duranti and Thibodeau to draw a distinction between the traditional record produced by a business process, and the Prospective Record demanded by the case study environments of InterPARES2*	
Rendering			
Re-usability / interoperability	Criterion	Extent to which a format is interoperable with software, services and tools enabling the content to be manipulated and reused for different purposes	
Robustness	Criterion	resistance to corruption	
Significant properties	Inspect: Wilson adapted	Characteristics of digital and intellectual objects that must be preserved over time in order to ensure the continued accessibility, usability and meaning of the objects and their capacity to be accepted as (~~evidence of~~) what they purport to be	
Simplicity	Criterion	Extent to which a format has no extraneous encoding ability beyond those required by the present Intellectual Object	

Stability	Criterion	Extent to which the format is managed to ensure orderly, predictable versioning with (usually backward) compatibility	
Standard		Published set of principles or specification with the aim of promoting good and / or common practice in a community	
Submission information package (SIP)	OAIS	An Information Package that is delivered to the producer to the OAIS for use in the construction of one or more AIPs	
Transformation	*OAIS*	*A Digital Migration in which there is an alteration to the Content Information or PDI of an Archival Information Package. For example, changing ASCII code to UNICODE in a text document being preserved is transformation*	
Transparency	Criterion	The readiness with which a format can be inspected or interrogated to discover its identity, content and attributes as against where these are obscured by compression, Wrapper formats, etc.	
XML eXtensible Mark-up Language	*W3c*	*Textual markup standard widely employed to promote data integration. Sometimes (not wholly accurately) asserted to be a file format*	
Verbosity	Criterion	Capacity demanded of storage media by Digital Objects / File Formats	
Vers Encapsulated Object (VEO)	Name	Data entity model developed and implemented in the Public Record Office of the Australian State of Victoria. At its simplest, VEOs employ the PDF file format for Content Data Objects and wrap in recursive layers of XML description	
Viability	Criterion	Closely related to Robustness	
Web2.0	*Wikipedia*	*Second generation of web development and design, that facilitates communication, secure information sharing, interoperability, and collaboration on the World Wide Web ..... the business revolution in the computer industry caused by the move to the Internet as a platform, and an attempt to understand the rules for success on that new platform*	
Wrapper format		A file format that encapsulates or wraps several bitstream that may themselves comprise instances of other file format specifications. May have the effect of reducing Transparency	
Xena	Name	Migration software produced and supported by the National Archives of Australia. Migrates common file formats to nearest OpenOffice.org equivalent, wrapping them and the original bitstream in XML	