Heritage of the People's Europe

HOPE
Heritage of the People's Europe
Grant agreement No. 250549

Deliverable 2.4

Best Practices for Trusted Digital Content Repositories

Version	2.0
Date	29/05/2012
Status	Final
Authors	Kathryn MATHE (KEE/OSA) Gabriella IVACS (KEE/OSA)
Dissemination Level	Public

HOPE is co-funded by the European Union through the ICT Policy Support Programme

Revision history

Version	Date	Authors / Contributors	Modifications
0.5	20/09/2010	**Authors:** Gabriella IVACS (KEE/OSA) Laszlo MAROSSY (KEE/OSA) **Contributors:** Kathryn MATHE (KEE/OSA) Jozsef BONE (KEE/OSA)	Milestone Report on the survey
0.75	14/06/2011	**Authors:** Kathryn MATHE (KEE/OSA) Gabriella IVACS (KEE/OSA) **Contributors:** Jozsef BONE (KEE/OSA) Joris JANSSENS (AMSAB) Titia VAN DER WERF (KNAW-IISG)	Best Practice Milestone Report
1.0	23/04/2012	**Authors:** Kathryn MATHE (KEE/OSA) Gabriella IVACS (KEE/OSA) **Contributors:** Lizzy KOMEN (Europeana) Jan MOLEJDNIK (Europeana) Valentine CHARLES (Europeana)	Deliverable Draft
2.0	29/05/2012	**Authors:** Kathryn MATHE (KEE/OSA) Gabriella IVACS (KEE/OSA) **Contributors:** Repke DE VRIES (KNAW-IISG) Mario MIELDIJK (KNAW-IISG) Lucien VAN WOUW (KNAW-IISG) Jerry DE VRIES (KNAW-IISG) Afelonne DOEK (KNAW-IISG) Donald WEBER (AMSAB) Eric BEVING (UPIP-BDIC) Hugo GUERREIRO (FMS) Armin STRAUBE (FES-Archive) Urs KAELIN (SSA) Jozsef BONE (KEE/OSA)	Deliverable Final

EXECUTIVE SUMMARY

The primary purpose of the HOPE Best Practice for Trusted Digital Repositories document is to guide HOPE social history partners, technical and content partners, current and future, in trusted digital repository practice. In this respect, the document looks past the short-term objectives of HOPE to a more sustainable federated repository model. To support this, it provides a thorough analysis of organizational attributes and local repository practice characteristic of the sector.

The introductory chapter defines "best practices for digital repositories" in the context of HOPE, and explains methodology used to gather and interpret qualitative and quantitative data. In addition to data collected from a survey given in the first months of HOPE, the T2.6 best practice team gathered together a resource of sample policies, workflows, strategy papers, and oral interview data. Such a resource enabled the team to ground its conclusions, but in the future, it could also serve as seed content for a knowledge base presented on the HOPE public wiki page.

Chapter 1, Framework(s) of the Social History Institution, depicts the rich landscape formed by today's social history institutions. To some extent, it follows the logic of Trusted Digital Repository (TDR) audit checklists and the Reference Model for an Open Archival Information System (OAIS), addressing cross-cutting issues, such as organizational viability, technological and financial sustainability, and procedural fitness. Case studies, written as free-standing pieces, highlight particular problems in the sector at the same time as giving detailed insight into its underlying praxis and ethos. The first section of Chapter 1 deals with Organizational Framework: Governance and Viability, the second one describes Legal Framework: Due Diligence, and finally the third section presents Technical Framework: Systems and Practices. Chapter 1 also provides the context for the HOPE model presented in Chapter 2. The audience for this chapter are HOPE technical and content partners and policy makers and funders interested in the sector as a whole.

Chapter 2 describes The HOPE Federated Repositories, comparing the technical architecture of HOPE to the OAIS functional model for federated repositories. The analysis is framed by a discussion of HOPE's policy infrastructure: its designated community, emerging content policy, and governance. The main body of the chapter details the HOPE architecture, component by component, in light of the model. Novel solutions implemented in the course of the project, like the HOPE PID Service and Persistent Identification or the Secure Storage in the Shared Object Repository, receive special attention as potential Best Practice for similar projects. The audience for this chapter are HOPE technical and content partners as well as like-minded federated projects or technical implementations.

Chapter 3 is dedicated to Managing Objects through Administrative Metadata and aims to define best practices beyond the state of the art of HOPE, guiding HOPE partners in the transition from access to preservation in order to ensure the system's viability over the

long term. Administrative Metadata is here viewed through the lens of the PREMIS standard and covers persistent identifiers, file naming, technical metadata, and fixity. The chapter serves to complement the HOPE Implementation Manual by offering the broader context for compliance. Intellectual Property Rights and Copyrights are not addressed as they are specifically covered by IPR Best Practice Guidelines. The audience for this chapter are HOPE technical and content partners and small- or medium-sized social history institutions.

The overall aim of this work remains to underscore the importance of reliability and trustworthiness for social history repositories and to guide the HOPE Best Practice Network in trusted digital repository management.

INTRODUCTION

Scope and Objectives of Task 2.6.2

- What are Best Practices for Trusted Digital Repositories? Are there best practices specific to social history digital repositories?

- Should best practices be prescriptive or descriptive? Should best practices present lessons already learned or a goal that has yet to be reached? The HOPE System as it is or as it could be?

- Should best practices unquestioningly accept the current state of the art—the models, standards, practices touted in professional literature? Or should best practices question these, move beyond, develop something to replace them?

Such questions have plagued us since we began our work two years ago on Best Practices for Trusted Digital Repositories in HOPE. Questions have come from all sides, sometimes from surprising quarters, and at times even from ourselves. In a sense, the following document is our response.

A Trusted Digital Repository (TDR) is one "whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future."[1] "In determining trustworthiness, one must look at the entire system in which the digital information is managed, including the organization running the repository: its governance; organizational structure and staffing; policies and procedures; financial fitness and sustainability; the contracts, licenses, and liabilities under which it must operate; and trusted inheritors of data, as applicable. Additionally, the digital object management practices, technological infrastructure, and data security in place must be reasonable and adequate to fulfill the mission and commitments of the repository."[2] A Trusted Digital Repository is generally considered to be compliant with the Open Archival Information System (OAIS) Reference Model.

Perhaps the primary dilemma we faced was that while Trusted Digital Repository literature and practice is about long-term viability, trustworthiness, and preservation, the HOPE Best Practice Network, as currently conceived, is not. It has the concrete short-term goal of standardizing and harmonizing metadata for inclusion into discovery services. It does not have the stated goal of developing sustainable digital object management. Nevertheless, HOPE is a best practice network, and as such must look beyond the short-term goals of a three-year project. While HOPE is not yet about the longue durée, it is about supporting access and seamless discovery-to-delivery for our designated community, and thus the network will sooner or later run up against problems

[1] RLG-OCLC, Trusted Digital Repositories: *Attributes and Responsibilities (RLG-OCLC Report)*, (Mountain View, Calif., RLG, May 2002), p.5.
[2] RLG-NARA Task Force, *Trustworthy Repositories Audit & Certification: Criteria and Checklist, Version 1.0*, (Dublin, Ohio, OCLC, February 2007), p.3.

of reliability and sustainability. Both Chapters 2 and 3 place HOPE's short-term goals in a long-term context.

It follows that Best Practices for Trusted Digital Repositories could not be completely descriptive. Neither the HOPE System itself nor, for the most part, HOPE Compliant Local Object Repositories can yet be called "trusted digital repositories" as defined. It was clear that external benchmarks would need to be formulated on issues such as digital object management, repository functions, and long-term data curation. Nevertheless, we hesitated to take a completely "top-down" approach. Instead, we looked closely at both the HOPE System and local institutional practice and attempted to identify "good practice" where we saw it. These can now be found in Case Studies and Chapter 2.

By the same token, it was not within the purview of our best practice work to overturn or otherwise challenge the "current thinking in the field." Still, we did attempt to ground the high-level abstractions of the OAIS model and generic terms of Trusted Digital Repository lists and PREMIS standard in the local experience and ongoing work of social history institutions in the HOPE federation. In so doing, we were able to place emphasis as we saw fit and at times to suggest possible lacuna or flaws in the models themselves. And while we were by no means revolutionary, the data and real life examples we provide may prove a useful check to those who carry out research in the upper echelons.

Perhaps the most difficult question is: what is specific about Best Practices for Trusted Digital Repositories in our sector? Small- and medium-sized social history institutional repositories struggle with sustainability issues, such as infrastructure, staffing, funding, and lack of forward-thinking policies as well as concerns about interoperability, data sharing, and broad-based access to collections. While these are concerns shared by many in the cultural heritage sector, we would argue that as private institutions highly engaged with contemporary users and issues, they have special risks and therefore special needs. By looking closely at the data gathered in the first phase of the project, we have tried to tease out these risks and address them in Chapter 1.

Methodology

Our best practice work proceeded according to the following steps. First, the current obligations, constraints, systems, and practices of all participating institutions were analyzed as part of the Content Provider Survey starting on June 28, 2010 and completed by August 1, 2010. The survey followed the logic of Trusted Digital Repository self-audit toolkits TRAC and Drambora, comprising questions in the following areas:

- Survey Section II/Questions12-28 – Policies
- Survey Section V.1/Questions 114-119 – Digitization
- Survey Section V.2/Questions 120-144 – Digital Object Repositories
- Survey Section V.3/Questions 145-197 – Digital Object Management
- Survey Section VII/ Questions 231-253 – IPR issues

Even though most of the surveys were turned in by the deadline, responses showed a varying degree of detail. In several cases qualitative answers were erroneous or left blank. In light of this, the best practices team decided to undertake follow-up phone interviews with all HOPE Content Providers (CPs). These interviews were conducted in

August and September, and the additional data highlighted the range of organizational and operational environments, underlying motivations and priorities, and differing interpretations and expectations of the HOPE Best Practice Network. Based on survey results and interviews, a set of Institutional Profiles was completed and made available on the HOPE internal wiki. In November 2010, the best practice team circulated an internal milestone Summary of Local Practices reporting findings and drawing attention to potential issues.

These results together with the HLD set the scope for the next phase. At the outset of year two, the best practice team focused on the major technical topics confronted by the WP5 SOR team as well as by local institutions developing repositories in house. An analysis of the state of current thinking and practice was carried out, and their relevance for HOPE assessed through the dual lens of PREMIS and the OAIS Reference Model. Topics presented in the Best Practices for Digital Content Repositories milestone report in June 2011 included: administrative metadata, technical metadata, fixity, file naming, and related workflows.

In the last phase, the best practice team has revisited our data on local institutional policy, infrastructure, and practice in order to analyze potential risks to reliability and trustworthiness. This analysis has been supplemented by case studies highlighting challenges faced by specific HOPE partners in the course of the project and solutions implemented locally—a sort of on-the-ground "best practice". The team has followed this with an analysis of the HOPE infrastructure as it currently stands, again through the lens of OAIS. Finally, the June 2011 technical milestone report has been extended with a section on permanent identifiers.

It should be remembered that many of our specific conclusions, particularly in the first chapter, are based on survey results and interviews and thus suffer from the problems inherent in these methods. Especially in the survey, technical terminology and broader language barriers caused confusion at times. Perhaps worse, the questions themselves sometimes reflected our own bias and the biases in the models we used, rather than the realities of the institutions we surveyed. On the other side, it was not always clear that the colleagues charged with responding to our questions, whether written or verbal, were in full knowledge of every aspect of institutional practice. Therefore, we have tried to base our conclusions on the aggregate of responses, the general drift, and the preponderance of evidence. When in doubt, we presented conclusions as speculative or conditional.

Another potential flaw in our method is timing. HOPE partners were surveyed and interviewed over the first six months of the project. It is quite possible and even probable that practice has since changed, whether in response to HOPE requirements or in the regular course of institutional development. We have tried to counter this with Case Studies, which give snapshots of recent developments at several local institutions. Still, we were faced with a grammatical dilemma. Should our analysis be presented in present or past tense. In the end, we opted to give the analysis in present tense for the following reasons: 1) perhaps most importantly, it simplified the written style and clarified the message; 2) an overwhelming majority of the results still hold true; and 3) the results reveal something about the sector as a whole, independently of HOPE. Simply put, we felt our results were strong enough to be presented as representative of this

sector at this juncture. Nevertheless, this does mean that in one or two cases past survey/interview results stand awkwardly next to recently completed case studies.
Each chapter is meant to stand on its own, more or less independently of the others in order to facilitate publication on the HOPE BP Wiki in the near future. Our work has been complemented by other best practice documents completed in the meantime: D1.3 IPR Best Practice Guidelines and MS12 Agreed Formats and Best Practices for Underlying Content, available on the HOPE internal wiki. We have also relied on other public and internal HOPE documents: D2.2 Common HOPE Metadata Structure including the Harmonization Specifications, MS5 HOPE Collection Policy Framework, MS3 Access and Use Conditions, the HOPE Implementation Guidelines, and the HOPE Glossary, the latter two published on the HOPE public wiki. When appropriate, these documents are referenced.

1. Framework(s) of the Social History Institution

1.1. Organizational Framework: Governance and Viability

Social history institutions, as defined within the context of the project, are those that collect material related to "the history of people's movements and individual life histories that were not part of official history, preserved by state archives and libraries"; that hold the "intellectual and material evidence of struggle and emancipation in written records, private papers, photographs, banners, posters, speech recordings and film". Such primary source material on mass movements and the everyday lives of ordinary people is rarely the focus of official archives and libraries—and also proves difficult to reach through traditional solicitation practices. Across Europe, organizations have emerged to fill this gap. Whether they stand alone or are affiliated with universities, political parties, or NGOs, these organizations have in common a strong thematic focus, active engagement in contemporary politics and social movements, and close ties with their community of users, be they scholars, journalists, activists, politicians, or public researchers. Keeping these at the forefront of their work, such organizations are often compelled to solicit material in unorthodox ways, to cross traditional professional domains, and to work on a project-by-project basis. Yet, their strong engagement with contemporary users and issues has spurred them to digitize and push content online more actively than many more traditional cultural heritage institutions. These organizations are often by necessity small, nimble, and reactive, and while this has allowed them to gather together a valuable corpus of previously overlooked source material and an active body of users, it has also hindered long-term planning and large investment into infrastructure or know how. Digitization when undertaken has tended to be *ad hoc*, and not supported by robust systems and workflows.

Common characteristics of social history institutions include the following:

- Mixed legal status, often affiliated to major universities, academic institutions
- Mixed funding, mostly private funding, strong motivation for fundraising
- Small- or medium-sized organizations, between 5-200 employees
- Shared ideology, strong political profile and commitment to organizations with similar mandate
- Research is integrated into library and archival activities, it constitutes an equally crucial part of their strategy
- Library, archive, and museum collection management practices are not sharply divided
- Strong international connections, already existing networks

The HOPE Best Practice Network comprises twelve institutions across ten countries and is a representative sampling of social history institutions in various organizational and socio-political contexts. All are members of the International Association of Labour History Institutions (IALHI). HOPE partners include:

- *Amsab-Instituut voor Sociale Geschiedenis, Ghent (Amsab-ISG):* Established in 1980, Amsab-ISG is a private archive, library, and museum and is officially recognized as a Flemish cultural heritage institution. The institution has been linked to Belgian socialist workers movements since its inception, though has recently broadened its range of activities. It has a subdivision in Antwerp, which is developing archival and library collections specific to that locale. Amsab-ISG is a member of the MovE – Musea Oost-Vlaanderen in Evolutie, a consortium of museums in East Flanders working to improve access to digital materials. Amsab-ISG has approximately 50 staff members.

- *Bibliothèque de Documentation Internationale Contemporaine and the Musée d'histoire contemporaine BDIC, Nanterres/Paris (BDIC):* BDIC is a state-funded library and museum run under the aegis of the French Ministry of Higher Education and Research. The museum and library function as separate departments under the same leadership and a common vision. Established in 1918, BDIC focuses its activities on contemporary history and international relations. BDIC operates as an inter-university research center for the Paris Universities and uses the infrastructure and technical services of the l'Université Paris Ouest Nanterres La Défense (formally Paris X). It includes records from its collections into SUDOC and CALAMES, both higher education union catalogues supported by the l'Agence Bibliographique de l'Enseignement Supérieur. BDIC is an officially associated unit ("pole associé") of the Bibliothèque nationale de France (BnF), coordinating acquisitions, preservation efforts, and digitization of materials. It is currently involved in several digitization projects and pan-European initiatives with the BnF, providing thematic content for Gallica and Europeana. It is likewise a member of CODHOS (the Collectif des centres de documentation en histoire ouvrière et sociale). There are 75 staff members in the two sections.

- *Confederazione Generale Italiana del Lavoro, Rome (CGIL):* CGIL Archive and Library is registered as a non-governmental organization which documents the activities of the CGIL trade union from 1944 to the present. Its sustainability is ensured by membership in a consortium that includes three major universities in Rome and the Consiglio Nazionale delle Ricerche (CNR). The consortium provides technical services and infrastructure. The institution has only four regular staff.

- *Friedrich-Ebert-Stiftung, Bonn (FES):* FES was founded in 1925. It is associated with the Sozialdemokratische Partei Deutschlands (SPD) and functions as a German political foundation receiving state funding according to the electoral success of the associated parties. It complements this with private funding, notably from the Deutsche Forschungsgemeinschaft (DFG). FES is engaged in a broad range of activities, most prominently political education (in a broad sense), funding of scholarships, and international cooperation. The Archiv der sozialen Demokratie (AdsD) and Library sections of FES maintain completely separate leadership and administration under the umbrella of FES (and are considered distinct partners in the HOPE project). They do share some facilities along with the services of a central IT unit, which hosts the website and oversees the back up of their systems. The library is a member of several German library

consortiums, while AdsD has plans to join a portal project run by the Bundesarchiv. FES has 600 employees.

- *Fundação Mário Soares, Lisbon (FMS):* FMS is a private foundation set up in 1991 to safeguard the political heritage of the former President and to spur the development of civil society across the Portuguese-speaking world. The archive and library sections were established in 1996 with the mission to use technology solutions to give access to their collections. FMS regularly provides expertise, technical support, and archival storage for digitization efforts in small organizations across the world.

- *Génériques, Paris (Génériques):* Founded in 1987, Génériques is a private association dedicated to preserving the history and memory of immigration in France and across Europe. Génériquesis a member of numerous associations related to immigration and documentation, notably CODHOS. There are a total of 18 staff members.

- *Internationaal Instituut voor Sociale Geschiedenis, Amsterdam (IISG):* Established in 1935, IISG has a dual legal status. The organization is primarily state funded; they are member of the Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), through which they obtain services and can also participate indirectly in Dutch and international projects. At the same time IISG has its own private foundation that handles some of its work under a separate budget. The institute specializes in the fields of social and economic history and is dedicated to preserving the legacy of social movements worldwide. IISG has 180 employees.

- *Maison des Science de l'Hommes de Dijon et le Centre Georges Chevrier, Dijon (MSH-Dijon):* MSH-Dijon and the Centre Georges Chevrier are both state funded research centers run under the aegis of l'Université de Bourgogne and the Centre national de la recherche scientifique (CNRS). MSH-Dijon is a federative structure which provides technical support to the eleven research centers in the social sciences and humanities at the university; Centre Georges Chevrier is one of them. The university provides technical services and infrastructure. MSH-Dijon is a member of CODHOS. There are 14 staff members at MSH-Dijon and 6 at Centre Georges Chevrier.

- *Open Society Archives at the Central European University, Budapest (OSA):* OSA is a private archive and research center established in1995 by George Soros to house the collections of the former Research Institute of the Radio Free Europe / Radio Liberty. The archive now collects records related to the Cold War and transition and global human rights. OSA also serves as the archive of the Open Society Foundations, which are active in over 40 countries. OSA is a unit of the Central European University and receives technical services and infrastructure from the university. OSA has 30 staff members.

- *Schweizerisches Sozialarchiv, Zürich (SSA):* Founded in 1906 to document the "social question" and promote access, SSA has dual status as a state-funded and private organization. Though much of its funding is provided through various levels of state and local government, it is run independently; additional funding

comes from a fee-based association. From the outset, the archive has been non-partisan, attempting to document all major political and religious tendencies in the region. Among its stated aims is to disseminate their materials through new technologies. It belongs to Nebis, the Network of Library and Information Centers in Switzerland. IT also provides AV content to Memoriav. There are 22 staff members.

- *Työväen Arkisto, Helsinki (TA):* TA was established in 1909 by the Finnish Socialdemokratiska Partis (SDP). Today it is a private archive maintained by the Labour Archives Foundation and regulated under the Finnish archival law. It receives subsidies prescribed by law and private financial donations for its operation. TA documents the history of the SDP and Finnish trade movement. The preservation of its digital masters is outsourced to the Mikkeli University of Applied Sciences. It cooperates closely with Työväenmuseo (the Finnish Labour Museum) to digitize materials. It is also planning to take part in the National Library of Finland's National Digital Library project, which will act as an aggregator for Europeana and support long-term preservation. TA has 10 employees.

- *Verein für Geschichte der Arbeiterbewegung, Vienna (VGA):* VGA was founded in 1959 as an independent "Verein" with the purpose to safeguard the intellectual heritage of the Austrian workers' movement and the so-called Old Party Archives of the Sozialdemokratischen Partei Österreichs (SPÖ). VGA is a non-governmental library, archive, and research center and is organized as a special branch of the Wiener Stadt- und Landesarchiv, who provide it with technical services and infrastructure.

As can be seen, HOPE partners or "content providers" (CPs) are, for the most part, private or non-profit organizations with independent legal status or only loose state controls; only three are completely state funded. While most stand as independent legal entities, they can nevertheless be highly dependent on the level of the national digital strategy (e.g. the Bundesarchiv portal, the Finnish National Digital Library, Gallica, or MovE), encompassed by state entities (e.g. BnF, Centre national de la recherche scientifique, KNAW, or the Wiener Stadt- und Landesarchiv), or part of the higher educational sector (e.g. Central European University, CNR, l'Université de Bourgogne, or l'Université Paris Ouest Nanterre La Défense). Only a few, such as FMS or Génériques, are truly stand-alone organizations. Most social history institutions consider external cooperation an essential part of their organizational mission, though such activities are not always well integrated into everyday operations. There are numerous forms of collaboration with a varying scope: union catalogues, cooperative digitization projects, shared infrastructure, and business partnerships to provide services or support. Local networks, strategic funding, long-standing commitments, and infrastructure generally take priority over international projects which often survive for a couple of years. Strategic partners and networks can form a community with almost as strong claims as an institution's target community of users. Though state funding is generally provided in limited measure, many also depend on funding from private resources and donations.

	AMSAB	Génériques	CGIL	FES (archive)	FES (library)	FMS	OSA-KEE	KNAW-IISG	TA	UPIP(BDIC)	UPIP(MSH)	VGA	SSA
mission statement	x		x	x	x	x	x	x	x				x
business succession or contingency plan						x			x				
collection policy	x	x		x	x	x	x	x	x	x		x	x
access policy	x	x	x	x	x	x	x	x	x	x		x	
privacy or data protection policy	x	x	x	x	x	x	x	x	x	x		x	
copyright policy	x	x		x	x		x	x	x	x		x	
reproduction (or download) service procedures	x		x	x	x	x	x	x	x	x		x	x
preservation plan	x				x	x			x	x			
back-up or duplication policy	x	x			x	x	x	x	x	x			x
hardware/software change or upgrade policy	x		x		x	x							
disaster and recovery plan	x		x		x	x							x
written digitization strategy or plan						x	x	x	?				x

1-A. Table - Policies & Procedures

Table 1-A lists policies and procedures addressed by the survey. The categories were established with the help of Trusted Digital Repository toolkits mentioned earlier: Drambora and TRAC. Almost all HOPE CPs possess a mission statement of some sort. Surprisingly, based on the survey responses it seems that the French partners and VGA do not. Disaster recovery, contingency planning, and other key documentation are also neglected according to survey data. Seen in one light, this data might highlight the difficulties in assessing organizational policy frameworks in the context of complex institutional arrangements. Such complexities are common at social history institutions, as many of these entities are integrated into state bodies or academic networks; custodianship is often undertaken at a higher level, and not always visible or well articulated at the level of the social history institution. The case of MSH-Dijon, a highly embedded organization, is illustrative in this regard.[3] However, seen in another light, missing policies might also reflect a lack of long-term vision and strategy. The fact that only two HOPE CPs were aware of business succession or contingency plan and five CPs of preservation plans does not bode well. Reviewing the list, it becomes clear that institutional policy frameworks focus on the here and now. Collection policies, access policies, privacy or data protection policies, copyright policies, reproduction service procedures, and back-up or duplication policies all scored highly. This is also the set of policies needed to run routine operations and user services on a daily basis. The long-term organizational viability is not addressed.

[3] MSH-Dijon confirmed in interviews that they have no written policies of the type listed in the survey. Their administrative policies and procedures are set at the university level. Their professional practice is guided by the French and international archival standards and guidelines. They follow the national guideline TGE Adonis for digitization.

Digitization plans or strategies were noticeably scarce, surprising given these institutions' clear focus on digital content (signaled by their participation in HOPE). Looking more closely at the possible types of digitization:

- large-scale systematic digitization, undertaken primarily for preservation;
- project-based or small-scale digitization, primarily for access;
- and on-demand or *ad hoc* digitization, generally as an internal or external reproduction service.

social history institutions strongly favor the short-term goals of small-scale projects. Eleven CPs listed access as a reason for digitization. Seven also listed preservation as a factor, yet as can be seen, only four had a documented digitization plan or strategy. Digitization priorities are, in fact, often set in the context of external commitments. MSH-Dijon are currently working on a digitization project with l'Institut National des Appellations d'Origine (INAO, the organization charged with regulating French agricultural products). IISG worked closely with the Koninklijke Bibliotheek to digitize brochures and other material. OSA cooperated with Columbia University's Butler Library to reunite a collection on Hungarian refugees after 1956. FMS are involved in several ongoing projects, including those with the Assembleia Nacional de Cabo Verde, the Arquivo Histórico Nacional de Cabo Verde, and the Instituto Nacional de Estudos e Pesquisa (INEP). And BDIC are regularly involved in thematic digitization projects with the BnF. Outsourced digitization is common. Many institutions mix in-house and outsourced digitization depending on the nature and time constraints of a particular project. Three HOPE CPs outsource all digitization activities, while seven mix in-house and outsourced digitization. With outsourcing comes risks; a loss of control over the quality and conditions of digital reproduction; and a potential loss of control over rights to material, especially with advanced digitization and enhancement techniques. Social history institutions may be unaware of the potential risks of outsourcing these activities.

More troubling may be the lack of concern over long-term system viability. While routine back ups are clearly taken seriously, HOPE CPs seem to lack a vision (or at least a documented vision) for their technical infrastructure and moreover seem to ignore the potential risk of disaster to their collections, facilities, and even staff. In this case as well, there is an increasing tendency to rely on external technical infrastructures for storing, backing up, and recovering digital content, again removing control and oversight from the institution itself. At least five HOPE CPs rely on external or supra- infrastructures for secure storage of digital masters and back-up of databases. It remains unclear whether institutions confirm that policies and procedures followed offsite meet institutional, community, and donor requirements, and if so, whether such arrangements are cemented in formal agreements. Securing and safeguarding content should be a first line of defense. And when it is a question of born-digital content or content that exists only in digital form, it is an imperative. Social history institutions often see it as their mission to protect rare private collections, or to rescue and secure endangered archives, or to safeguard politically sensitive material. Back-up policies and levels of disaster preparedness are, in fact, closely linked to this mission, having a real impact on institutional reliability among many other things.

All this raises questions as to whether the business models currently followed by social history institutions are sufficient to support strategic planning in a more comprehensive

manner. Or to be more precise, whether the development of digital services would fit into the current business model. During interviews, CPs complained of a continuous dilemma over where to invest resources and how to prioritize between analogue and digital services. Many also confirmed that digital services are still viewed as an extension of their analog counterpart, and therefore play a less integral role in daily operations. New systems seem to exist in uncomfortable parallel with legacy systems. This constant balancing between "types of services" reveals an *ad hoc* approach as well as a lack of strategic thinking.

On the one hand, social history institutions with an aim to serve their publics and fulfill their non-public mandate, are feeling pressure from the growing number of users who have moved online and who have high expectations to be served online. Institutions feel compelled to offer up content and services on social sites, such as Flickr, Facebook, and Scribd, as well as on cultural heritage portals like Europeana. It is thus not surprising that several HOPE CPs have turned out to be early adopters of new user-centered technologies and services. AdsD ran a crowd sourcing project, uploading selections from their photo collection to Wikimedia under a Creative Commons non-commercial, no derivatives license. OSA developed the Parallel Archive, a collaborative digital humanities tool allowing users to upload and form communities around archival sources. Génériques created thematic blogs "Melting Post" and "Générations" to explore topics of interest using sources from their collection. On their "Hoje no século XX", FMS provide dynamic feeds of newspaper headlines from throughout the last century. On the other hand, social history institutions are increasingly compelled to compete for funding subsidies to support their work, without clear evidence that temporary non-structural funding can sustain their growth online. The situation may in fact create an incentive for experimentation with new entrepreneurial approaches to digital services in order to recover some of their costs. IISG may be a forerunner of the next wave with their Social History Shop, which combines the best features of a library search engine and an online retailer to offer up high quality reproductions of original archival content.

To help CPs address these issues, HOPE recommends the use of a Trusted Digital Repository self-audit toolkit, such as TRAC, DRAMBORA Toolkit, Nestor, the Data Seal of Approval, or the Data Asset Framework (DAF). The ultimate goal of this effort should be that the institution can: 1) articulate and document its own missions, aims and objectives, shortcomings and potentials; 2) inventory its activities and assets; and 3) be aware of pertinent risks and try to resolve these. Social history institutions, like many other cultural heritage institutions, are in the transitional phase between traditional collection management and digital object management. Yet taken together, the digitization of analog collections, storage and management of content, and expanded access do not form a break with former practice but an extension of it. These institutions will continue to serve their designated communities in a traditional way, offering records physically in their reading rooms or on offline networks to comply with legal regulations. It is also true that digitization and digital curation is a costly endeavor and not all analog collections need to be digitized. Therefore prioritization and planning remain an issue. The audit process would allow these institutions to analyze their strengths and weaknesses and to respond in a systematic fashion as part of their high-level business planning.

Case Study: Strategic Planning at the Internationaal Instituut voor Sociale Geschiedenis (IISG)

The Internationaal Instituut voor Sociale Geschiedenis (IISG) was established in 1935, and currently functions as an archive, library, and research institute working under the aegis of the Koninklijke Nederlandse Akademie van Wetenschappen (KNAW). Originally founded to house the Netherlands Economic History Archive as well as to safeguard labour union materials threatened by the rising tides of war, the IISG now house extensive collections in the field of social and economic history and the documentation of social movements across the globe. IISG currently hold more than 3000 archival fonds or collections, including inter alia some 250,000 photographs and 120,000 posters, as well as a substantial library collection.
With more than 180 employees, the IISG are one of the bigger institutions represented in HOPE and also coordinate the project. Needless to say, IISG hope to benefit from the range of opportunities presented by the HOPE project and are leading the development on the SOR shared storage system. They also maintain the HOPE PID Service.

IISG stand alone as the only HOPE partner to have developed a comprehensive information strategy. The IISG's Information Policy Plan formulates the goals and underlying policy principles of digitization and computerization at the institute. It also sets out the framework for further development of the institution's information systems, a few of which are now due for replacement. To complement its plan, IISG also performed self-audit using the DRAMBORA Toolkit in order to assess risks management issues and preservation planning needs. Both the assessment and the plan have provided a considerable knowledge base for the organization.

According to the plan, IISG should aim at using standardized, generic technologies to ensure interoperability for broadly defined "data sharing". Open standards and open ICT architectures are mentioned as relevant in this respect. The general aim is in line with IISG's vision to lead "research in the field of 'global labor and economic history'", which as envisioned would be achieved through analysis of large quantities of comparable historical data. The plan advocates integrated work flows, the coherent management of information assets, a wholesale presentation of collections online, seamless access to cultural heritage resources, the introduction of novel research tools, the enrichment of metadata through data mining tools and techniques, and finally the creation of an infrastructure suitable for long-term storage and access. IISG's ambitious strategy would nicely underpin the networked research activities currently advocated by scholars of digital humanities.

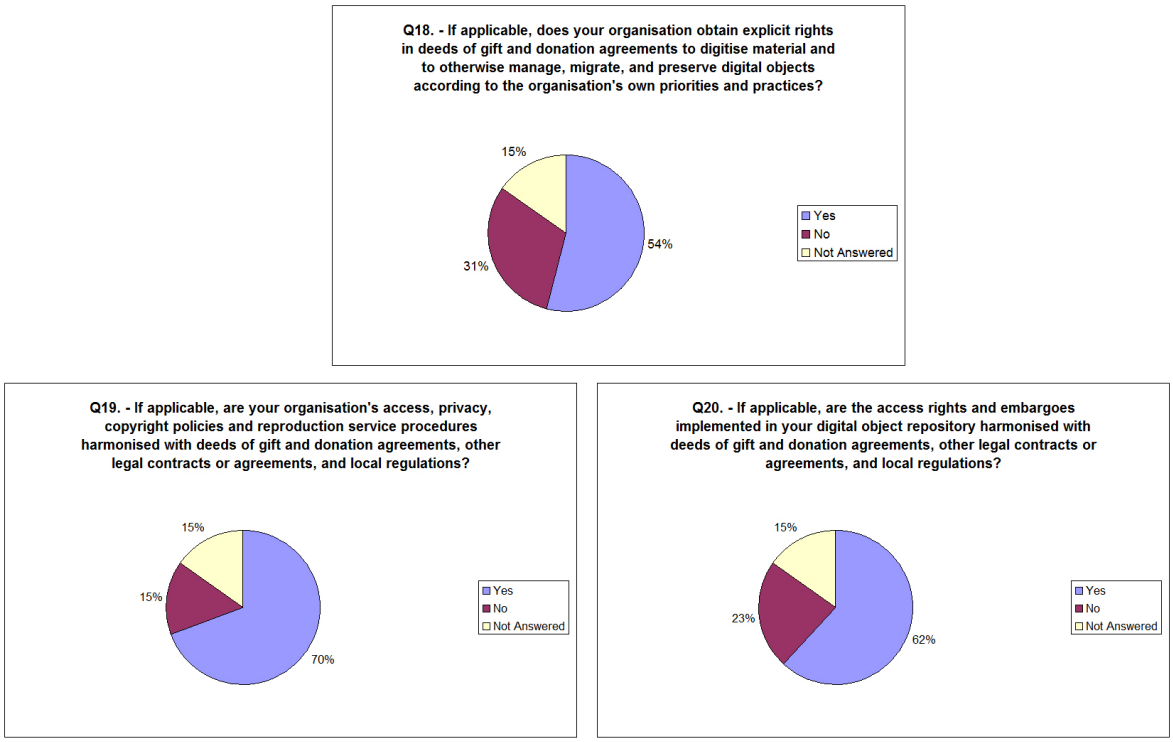
In its plan for 2008-2012, IISG have set the objective to conform to standards and best practices in the sector. IISG also restructured its operations by setting up a dedicated digital service unit as a central organ in the organization. Without detailing related institutional policies, the plan likewise demonstrates that ICT planning is not a stand alone process. It is dependent on other policies and objectives: a clearly defined designated community; a research agenda—a crucial element in the mission of all social history archives—; and an integrated organizational framework. The plan also frames IISG's activities within a wider network of like-minded institutions. IISG already have strong ties with KNAW, but DANS, the SURF foundation, and DRIVER are also presented as important collaborative partners, as is IALHI, the network behind the HOPE project.

IISG are working to build a comprehensive infrastructure that is suitable for the long-term storage of and access to digital collections on their way to implementing a Trusted Digital Repository. The IISG Information Policy Plan sets a clear-cut agenda and provides a template that is readily exportable and adaptable for all social history institutions.

1.2. Legal Framework: Due Diligence

The sheer complexity of currently held analog and digital material and the several workflows for managing this material present a real challenge for legal compliance. The holdings of a given social history institution may cover a range of content formats, among these "archives, books, periodicals, brochures, leaflets and pamphlets, visual documents such as posters, prints, cartoons and photographs, audiovisual and sound recordings, banners and paraphernalia." And digital material is no less varied—whether it be digitally-born or digitized content: a simple image file scanned in the 1990s, a

database from the 1980s, or complex multimedia object created very recently; a proprietary or open format; a preservation-quality master object, a watermarked access copy, or a thumbnail. Social history institutions working amidst peculiar conditions to digitize unique collections or to accession externally digitized materials have little control over formats and quality; however, they should always be in control of statutory provisions, deed of gifts, donation agreements, and licenses to eliminate ambiguity in the ownership and reuse of collections.



1-A. Diagram - Policies

As seen in the previous section, most institutions have a developed set of in-house legal policies or have adapted policies set by major cultural heritage entities. Eleven HOPE CPs have data protection or privacy policies, while nine have copyright policies. Still, the status of digital materials for both short-term and long-term access often remains unclear. When CPs were asked whether "they know about digital donations" and whether "the current policy framework treats sufficiently the content existing only in digital format", several admitted to being involved on projects when digital content was given away with no legal provisions or when another organization donated only digital copies to them. In other responses, quite a few CPs suggested that digital records were merely copies of a physical records, and as such deserved no special status. As shown in diagram 1-A, only 54 percent obtain specific rights to digitize material and manage digital copies. Their policies focused completely on the management of analog material and copies of analog material (although IISG, FMS, SSA, and OSA have already started accessioning digitally born materials). Both the survey analysis and numerous contradicting statements during the interviews proved that legal compliance on both digitized and born-digital content as well as related preservation strategies, are not

addressed in a comprehensive manner at many partner institutions. Only 62 percent claim to harmonize the rights they obtain over digital content with access (23 percent state that these are not harmonized).

It has, in fact, become crucial to expand the practice of "due diligence" beyond our physical collections. The following problem areas have been identified:

- Donation agreements, deeds of gift, and deposit agreements establish a legal relationship between the donor and the donee. The donor often owns the rights over the physical object as property but does not own the intellectual property rights. Donor restrictions can often overwrite national legislation in private institutions; also, they can add additional embargo time on access.

- Digitization for preservation needs to be clearly distinguished from digitization for distribution and public use. On the other hand, large scale digitization, even with the goal of preservation, involving third parties in the process could raise copyright concerns.

- Social history archives and libraries need to obtain the rights to transform material for preservation purposes. Responsibility for preservation has traditionally been considered as attendant on ownership of analog files; ownership of digital materials is a less straightforward matter, as digital materials are less tangible.

- Rights to distribution are provided through licensing arrangements; there is a lack of standard licensing models for non-state organizations.

- Clearing rights on copyrighted and orphan works puts an extra burden on social history institutions due to their limited resources and broad collection scope. Exception-based copyright legislation cannot be effectively applied over collection-based solicitation; each collection consists of different types of materials from different copyright holders.

- Social history collections by their very nature include sensitive data about private people; data protection procedures require a tremendous amount of staff time and financial input to comply with.

The legal barriers to access and reuse of material remain high. A majority of HOPE partners list IPR as an obstacle to their work. Six CPs explicitly remark the presence of orphan works in their collections, often in large quantities. For those whom IPR is less of a problem, third party privacy has been an issue. One CP also mentioned "secrecy" laws in the case of state documents. As noted above, the connection between rights and access is tenuous and heavily dependent on informal practices or manual intervention. HOPE CPs confirmed that copyrights and other legal restrictions are the main basis for restricting access to digital and analog content. Yet copyright and other restriction information is stored in local systems as loosely-controlled free-text metadata—a practice which served for physical materials but does not extend well to online publication. Access is generally controlled at a high level (fonds, series, record group) through informal means. In many cases, restrictions are reflected in digitization priorities, as restricted

collections are not digitized in the first place. In the most extreme cases, digital materials are available onsite only. HOPE partners limit the reuse of digital content presented online primarily through physical barriers, such as the provision of lower-quality derivatives and watermarked copies. Licenses or legal clauses stipulating permitted uses are offered as part of reproduction services, but not for material presented online. With such informal practices, it can be surmised that if in doubt, or simply for convenience, institutions rather err in the direction of over-restricting their collections. It is telling that Amsab-ISG, OSA, and MSH-Dijon are the only three CPs who mention institutional opt-out policies.

As a rule, social history institutions disagree with the current exception-based copyright regime, which does not take into consideration the unique value of historical collections as a whole and their importance for research and educational use. These institutions consider their activities—providing both short- and long-term access—a public and non-commercial service. There is a general consensus that Notice and Takedown Policies, in other words an opt-out model, would be more appropriate. In this case, out-of-print publications, orphan works, or works of unknown copyright status could be put online. When copyright clearance is possible, institutions should make a best effort to clear the rights, suggesting several licensing options to copyright holders, among them Creative Commons licenses or no limitation at all. HOPE recommends that institutions clarify the permitted re-use of online materials prior to works being put online, and to include them in the donation agreements, which should be harmonized with licensing regimes. Rights management should be folded into routine accession and processing/cataloguing workflows and machine actionable data be captured in collection management and digital repository systems at a high level of granularity. Rights management should cover digital and physical copies without discrimination.

What remains to be tackled is access and curation in the longer term. Professional attitudes towards digital curation and digital repository management must change; digital materials should be understood as more than copies of physical collections and an integrated approach to material in its multiple formats should be formulated. Most social history institutions have strong opinions about access to their collections in the short term with some ideas for addressing it. HOPE CPs have shown themselves willing to comply with the legal requirements set by the EC and Europeana, carefully selecting openly accessible collections and assigning Europeana licenses to digital materials. However, social history institutions must also take care to secure the rights to manage material, in whatever format, over the long term.

Case Study: Compliance at the Schweizerische Sozialarchiv (SSA)

The Schweizerische Sozialarchiv (SSA) was founded in 1906 in Zürich (Switzerland) to document the "social question" and promote access to collections. From the outset, the archive have been non-partisan, attempting to document all major political and religious tendencies. Collections currently reflect themes such as: gender and age relations, migration, labor history and trade unionism, social policy, political parties and social movements, the environment, and communication and transportation. Holdings comprise over 27,000 units (2.5 kilometers) of archival material, including more than 100,000 photos, posters, and visual objects; 150,000 books; and press and propaganda material numbering almost 2 million items. One of their main stated aims is to employ new information technologies and new means of dissemination to give access to their materials.

SSA joined the HOPE project as an external partner to support their mission of dissemination and access through technology. They are also participating in the development of HOPE's shared storage solution and see the experience as valuable to their own digital object management work.

Before starting the HOPE Project, SSA took the precaution of ensuring that formalized agreements were in place for all donations from corporate bodies and private individuals. All agreements include clauses treating: access conditions for donors and depositors, access rules on the reuse of the documents by third parties, and copyright issues. As a standard practice, depositors or donors give SSA authorization to photocopy, microfilm, or scan documents for non-commercial purposes; they are likewise encouraged to permit online distribution of documents via the institutional web site. About 95 percent of all archival holdings currently held by Swiss Social Archives are accessible without any restriction. Only a few collections, and selected record groups and series, are subject to access restrictions. Such cases include personality-related records from the gay and lesbian movement, copies of state security files, and holdings on extremist parties.
Documents with pornographic, racist, or sexist content are also restricted based on in-house policy. In addition to physical material and audiovisual media of every kind, SSA accepts digitally born records along the same contractual line.

The question of copyright and data protection has always been treated by the Swiss Social Archives in a very pragmatic way. They are the only institutions to lack formal policies on Access, Privacy/Data Protection, and Copyright. They have, moreover, deliberately taken some risks, when the importance of access to the material clearly outweighed the collective or individual interests to be protected. As the result of this practice, only a few conflicts have developed and all disputes were settled out of court. The online presentation of digital objects and metadata has considerably changed the whole setting.

The HOPE project was an important impetus to review IPR issues from a different perspective: copyright, access, and reuse had to be fixed and, where possible, to be conclusively resolved. SSA management noted from the beginning that Swiss legislation differed in several respects from copyright laws in other countries; compared to other European countries or the US, the Swiss Copyright Law is more permissive about infringement. For example, under Swiss federal law there is no absolute prohibition of circumvention and the download of content from the internet for private use is permitted free of charge. As part of revising in-house access rules, data protection was strongly taken into account, and online material was reassessed potentially sensitive content. As a result few files, including audio recordings of the meetings of various trade union bodies, have been blocked or deleted from the institutional site.

The SSA experience reveals how due diligence can affect everyday practice. Rights obtained early in the archival workflow have given them the freedom to broadly disseminate their holdings. The case also illustrates the effect of national legislation on local institutional policies and practice.

1.3. Technical Framework: Systems and Practices

The introduction of networked systems in social history institutions dates back to the 1990s. In the wake of large-scale library automation, such institutions were eager to convert their card indexing systems or paper-based finding aids into electronic catalogues and to this end introduced specialist or standard softwares. In addition to collection management systems, other databases were gradually brought in to manage information about items of a single media-type, collection, project, or exhibition, particularly in the case of archival collections—for which item-level data has traditionally been scarce—and visual collections—for which descriptive standards are less widely applied.[4] Such shadow catalogues were presented on institutional websites alongside

[4] Here it is notable that almost 50 percent of digital items slated for submission to HOPE were described with idiosyncratic descriptive systems. Not surprisingly, archival collections were in a much worse situation than library materials as institutions struggled to describe digitized material at a high level of granularity;

standard catalogue entries. In some cases, another layer has recently been added, as systems were brought in to manage a burgeoning supply of digitized content. In other cases, digital content sits on file servers under loose controls, from where it is pulled into descriptive databases or pushed directly to websites. These self-made, patch-work type information systems are now ubiquitous in the cultural heritage domain, and even more so in social history institutions where, with their relatively small-scale and heterogeneous collections, powerful enterprise management softwares have failed to take off.

The legacy of data structures and systems in use at social history institutions has not only led to an outdated and often expensive information architecture but also at times obstructed the introduction of best practices. Despite the widespread acceptance of library, archival, and museum descriptive standards (and respective XML schema) and emerging importance of preservation standards, legacy systems are often not easy to adapt. Those institutions depending on in-house or open source solutions often lack the technical and professional know-how to keep abreast of changing practice. Those using proprietary solutions are locked into data structures supported by the service provider. In all cases, strong institutional habits bind these organizations firmly to existing practice—no matter how outdated. It is therefore not surprising that none of the HOPE partner institutions have managed to implement a fully-functional preservation repository along the lines of the OAIS model. (In fact, one of the clearest conclusions that can be drawn from the survey data is a general lack of consensus over the concept of "digital repository".) On a broad scale, but also within individual institutions, an interesting mix of proprietary, open source, and custom built systems co-exist. Manual processing and workarounds are often used to compensate for true integrated system architecture.

Only three institutions have full-scale digital repositories: FMS's Westbrook Fortis, SSA's IMS Server-Client, and VGA's M-Box. All three are proprietary solutions to some extent. FMS manage their development in house, while SSA and VGA depend on services provider to maintain and develop their systems. While none of the three systems supports full preservation functions, all include some ingest, storage, and access functions (validation of size and/or formats, fixity checks, derivative creation, access controls, and collection and storage of technical, structural, and provenance metadata). More surprising perhaps, all support descriptive metadata internal to the system, rather than linking to existing collection management systems—as all three institutions are archives, granular metadata may not be available elsewhere. The fact that all three solutions are proprietary may limit their ability interoperate with external services. Currently, none of them do. In the case of Westbrook Fortis, data is also "locked in" to a proprietary file format.

Other institutions such as IISG, FES Library, and OSA have experimented with open source digital repository software, Fedora, MyCoRe and D-Space respectively, for special projects and non-HOPE collections—in the case OSA and IISG to handle born-digital content. In all cases, the software have been configured to handle some ingest, archival storage, and data management functions. Interestingly, none of these institutions has yet developed these solutions to fit their entire collections. The majority of HOPE partners

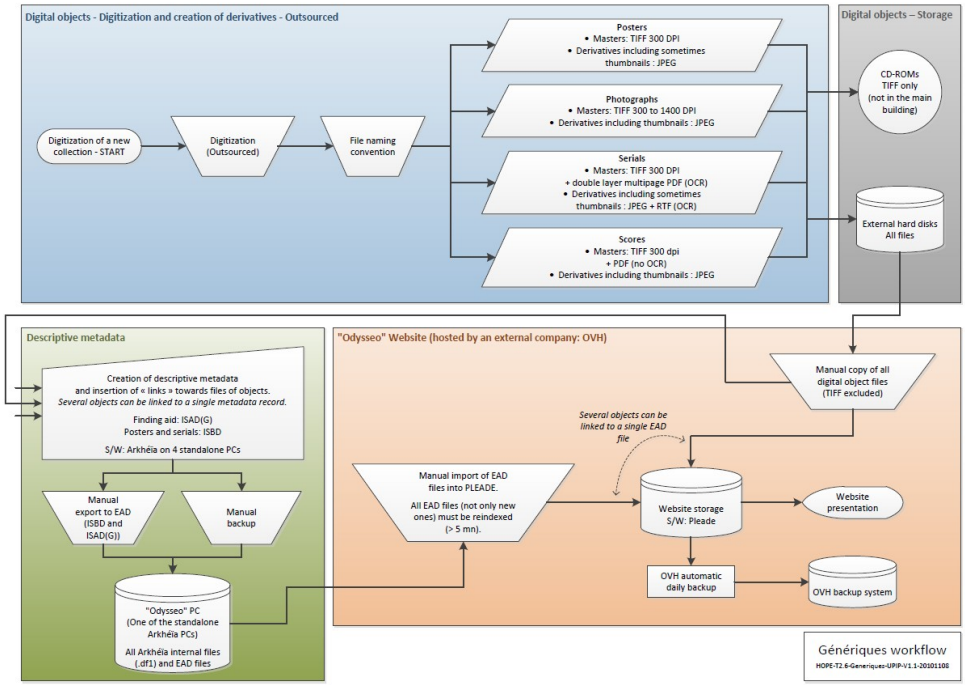
approximately 85 percent of archival items were described using idiosyncratic systems. Visual and museum materials were also disproportionately described using homegrown systems; not a single dedicated museum or visual standards was used by HOPE CPs.

continue to store digital content on file servers or in collection management systems. To the question "If no digital object repository is currently in place at your organization, how do you store and manage your digital content?", responses were more or less variations on the same theme. Six CPs explicitly mentioned storing content on file servers; of these three mentioned links from metadata records. Two store digital content directly in its collection management system. Seven CPs routinely back up digital masters to tape, storage devices, or servers; several depend on a larger umbrella organization to perform this task.

Three CPs indicated that they have service level agreements with external providers, but more use proprietary collection management systems for at least a sub-set of their material: Adlib, Aleph, Alexandrie, Arkhéïa, Flora, FAUST, Geac are used singly or in some combination by five institutions. Such solutions may hinder the standardization of metadata across partner institutions and obstruct effective bridging with external services. Three CPs use open source solutions supported by local governments and professional associations. CGIL use the Italian implementation of UNESCO's CDS-ISIS, called CDS-ISIS Teca. FES (Library) use Allegro-c developed and supported by the Science and Culture Ministry of Lower Saxony and used widely throughout Germany. SSA use Nebis, a Swiss library union catalogue. Two CPs have recently introduced international open-source library platforms, Greenstone and KOHA. Such open source solutions exhibit more flexibility in their service packages than proprietary solutions, though integration with other systems would still require staff time, expertise, and possibly commitment to the community development process.

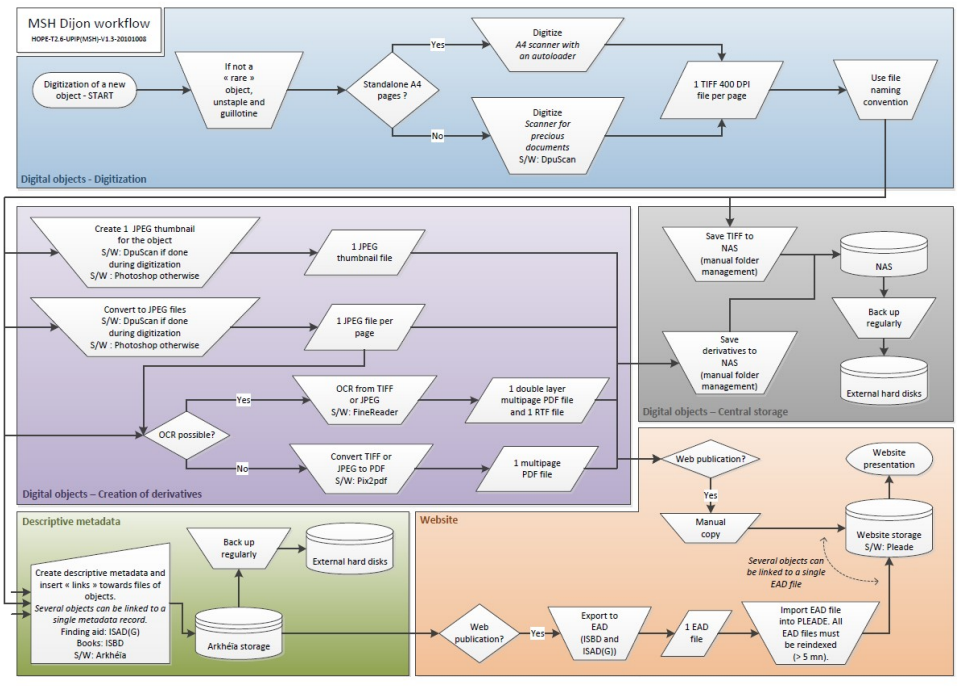
Turning to high-level digital object management workflows, it is clear that likely as a result of limited financial resources and staffing, many of the smaller organizations rely heavily on service providers or umbrella organizations for IT support and infrastructure and many likewise outsource digitization. Not surprisingly, institutions with more service dependencies tend to have more straightforward internal workflows, containing fewer loops, redundancies, and extra manual work in the process itself. Larger institutions and those with a more varied collection profile tend to have more idiosyncratic, flexible, and "organically" developed internal workflows. In these cases, scanning is mostly run by in-house staff with in-house ICT support, while large scale digitization is outsourced to vendors. In contrast, institutions with digital repositories tend to have more standardized digitization and more uniform workflows.

The following diagrams show the high-level digital object management workflows of the three French CPs.



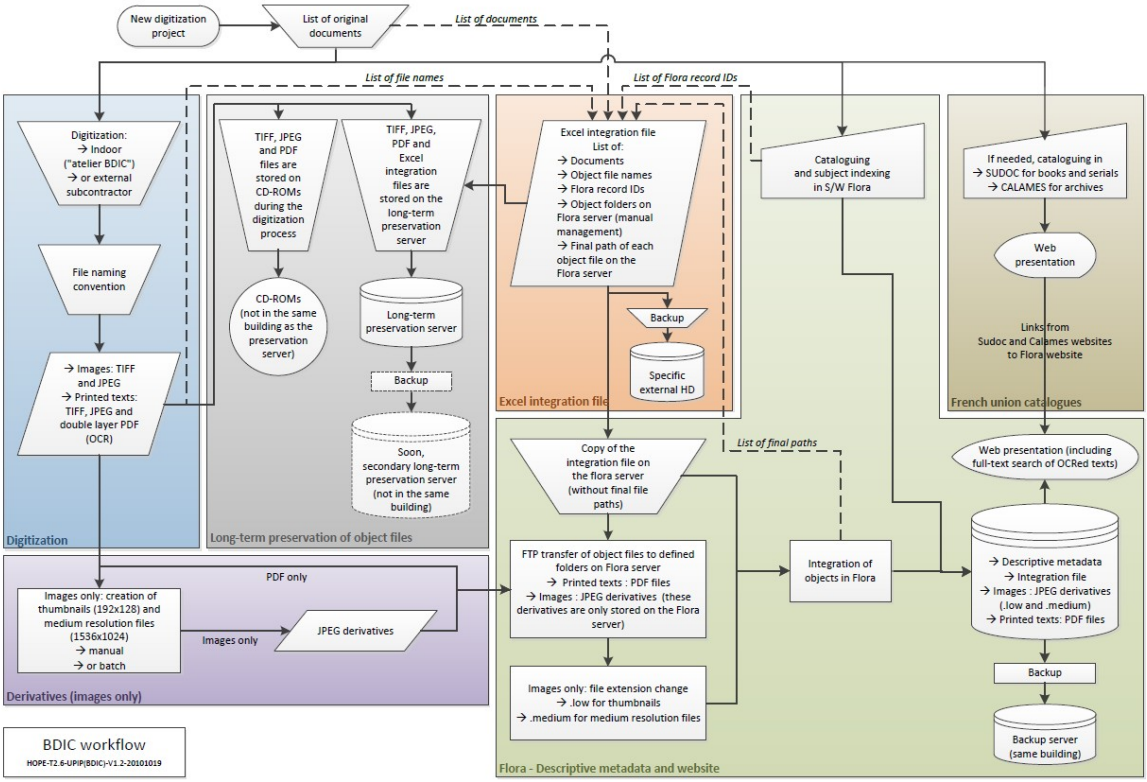
1-B. Diagram – Génériques Workflow

Génériques: digitization is outsourced; description is managed through the proprietary software Arkhéïa with indexing and web publication through an open source extension Pleade; internal workflows are highly dependent on manual procedures.



1-C. Diagram – MSH-Dijon Workflow

MSH-Dijon: digitization is managed in house, complicating workflows; description, indexing, and web production are also based on Arkhéïa/Pleade; the university infrastructure means fewer manual procedures and more robust storage and back up.



1-D. Diagram – BDIC Workflow

BDIC: digitization is both internal and external; collection management, indexing, and web publication are managed through the integrated solution Flora, thus cutting steps; links from the French union catalogues, SUDOC and Calames, require synchronization with Flora; university infrastructure provides robust storage and data back up; work is focused around an excel-based "integration file" to manage disparate activities.

Given the understandable investment and attachment to legacy systems, HOPE can only give a gentle nudge towards best practice. As a first step, institutions should become familiar with the functional entities (in OAIS terms) of a preservation repository system: ingest, archival storage, data management, and access, as well as the processes that each of these include. By articulating more clearly the entire range of functions, an institution may come to a better understanding of what a digital repository is and what it isn't. It is recommended that an institution prioritize functions according to current need and available resources and begin to gradually introduce functions into local architectures. When developing a repository system from scratch, HOPE recommends a loosely-coupled, modular set of components, whether packaged as a single system or a stack of applications. In general, HOPE advises the use of open source solutions as the strongest protection against data lock-in in its various forms. Currently, the open source solution Fedora (Flexible Extensible Digital Object and Repository Architecture) nicely

fulfills the above requirements and is put forward for those with the technical capacity to develop and support it. Those who lack the in-house technical expertise should consider outsourcing the development of Fedora or another open source package; this may not cost more than a typical service package. In general, institutions are advised to avoid monolithic one-size-fits-all solutions, but should focus instead on forging a system of different elements—a hybrid of open source and proprietary components if necessary. Finally, for small institutions that lack funding and technical know how, distributed or federated services shared among like-minded institutions may be a good alternative to profit-driven service providers.

As it stands, intentions regarding the use of open source technologies are clear: the survey revealed a strong preference for open source applications and open formats. In practice, many CPs explicitly committed to using open source technologies to develop digital repositories, OSA, IISG, Amsab-ISG, are still in the research and pilot phase. Those who have actually set up digital repositories have opted to use proprietary systems. This may be a worrying trend, and HOPE recommends that institutions take a longer look at open source alternatives and applications with modular and loosely-coupled architecture. More problematic may be the fact that institutions have clearly articulated the need for dedicated repositories as separate from collection management systems. HOPE CPs have either a collection management system or a digital repository but not both. For the most part, collection management systems, many proprietary but some open source, remain the focal point of institutional workflows and technical development. Until institutions begin to clearly distinguish digital object management from their more familiar collection management, Trusted Digital Repository best practices may remain elusive.

Case Study: Adapting Practice at the Archiv der sozialen Demokratie (AdsD) of the Friedrich-Ebert-Stiftung (FES)

Established over 40 years ago, the Archiv der sozialen Demokratie (AdsD) in Bonn, Germany is the archive of the Friedrich-Ebert-Stiftung (FES). FES is associated with the Sozialdemokratische Partei Deutschlands (SPD) and functions as one of several German political foundations. Alongside the extensive holdings of the SPD, AdsD hold records from organizations and people prominent in the German labour movement as well as more recent collections related to the peace, environmental, and women's movements. They also have an extensive audio-visual collection, including approximately 1.2 million photos, 67,000 posters, 50,000 pamphlets, 250 historical banners, and 22,000 film, video, and sound documents. There are more than 600 employees at FES, approximately 60 of whom work at AdsD. AdsD have used the proprietary archival management system FAUST for 20 years; FAUST does not support a particular descriptive standard but allows curators to create collection-specific templates. For AdsD, HOPE was an opportunity to give greater access to their holdings through Europeana and the IALHI portal. The HOPE Best Practice Network has also helped guide them in the standardization of their disparate metadata sets. They are not using the SOR or the HOPE PID Service.

HOPE requires content providers to map descriptive metadata to the common HOPE schema and to provide actionable PIDs resolving to a digital object and a "landing page" providing context for the object. AdsD found itself unable to meet these requirements for two reasons. First, they were unable to provide a direct link to every single object and its description via the existing system. Second, descriptive metadata in FAUST had not been harmonized internally or with external standards; as the system had been used over many years by numerous staff, a daunting number of descriptive templates had accumulated.

FAUST was already set up to provide access to metadata via the internet, but to meet the requirements of HOPE, the solution needed to be upgraded. A task force was set up to implement a solution for this web access and succeeded on this account.

To facilitate the export of the HOPE collections, a new simplified database was developed in FAUST; the metadata template was designed to capture metadata from all previous templates in a standard form that was easily mapped to the HOPE Visual Profile.
Transformation, correction and standardization of the descriptive metadata took a great deal of time and effort, but the new database now successfully integrates all collections slated for submission to HOPE as well as several others. Essential to the success of the endeavor was transparency with all parties concerned. Staff generally reacted positively to the challenge, since the reorganization proved an opportunity to address a number of problems with the databases.
Perhaps more importantly, the benefits of describing material according to standards became apparent. As a result, the AdsD is able to export its metadata in good quality to HOPE.

As it stands, AdsD are still in a transitional phase of development. Though they now have a unified descriptive system, no dedicated digital object repository has yet been set up. Master files and derivatives are still uploaded to a file server which is backed up regularly. The only connection between FAUST and the file server are the document signatures, which make up part of the file names and are recorded as part of the descriptive metadata in FAUST. Thumbnails continue to be stored directly in FAUST and can be provided for HOPE aggregation. In the future it is possible that FAUST itself can become a central component in a full-scale digital repository. AdsD analyzed FAUST for OAIS compliance (see: AdsD newsletter 2008) and ran a pilot on three digitally-born collections. FAUST proved adaptable to storing technical metadata, which could be promising.

In the case of AdsD, HOPE provided the impetus to change long-standing institutional practice. Internal consensus played a key factor in the success of their endeavors. The case also reveals the power that the collection management system, as such, exerts over institutional digital object management practices.

Case Study: Digital Object Management Reborn at the Fundação Mário Soares (FMS)

The Fundação Mário Soares (FMS) was founded in 1991 in Lisbon (Portugal) to carry on the legacy of Portuguese President and socialist politician, Mário Soares. In the spirit of Mário Soares, the private foundation caters to a broad and diverse public, seeking to foster the free debate of ideas and values and an engaged civil society. The archives were set up in 1996 initially based on the personal papers of Mário Soares; they continue to collect around issues of relevance to the contemporary Portuguese and Portuguese-speaking worlds, including many holdings from the former Portuguese colonies as well as a rich photographic archive. FMS depend completely on their internal IT unit to maintain and develop their systems.

Since their inception, FMS archive have had in place a mass digitization policy, whereby digitization has been integrated with physical collection processing and description. All processed records are digitized, and most are made directly available on local stations at the archive itself. FMS have used the proprietary software Westbrook Fortis to support their activities. Fortis supports digitization, ingest, storage, description, and access. But it also stores content in a proprietary format, which has obstructed their efforts. FMS are heavily involved in external digitization projects, cooperating with small organizations and NGOs around the world to provide expertise, resources, and safe storage for their archival materials. Interestingly, as a result of their early adoption, FMS are also the first of the HOPE partners to suffer from obsolete or outdated digital content on a large scale, and slowly they are being compelled to re-digitize or migrate earlier collections. FMS see the HOPE project as a means for providing access to their rich store of digital content. They also view HOPE trusted repository best practices as useful input to their long-standing digitization program. FMS are not using the SOR but plan to use the HOPE PID Service.

By the time FMS joined the HOPE project, they were reaching the limits of what they could do with Fortis. A lot of time and effort were being put into making it work with different media types, and the system was working against their efforts to make collections available online in a sustainable way. In that sense, the HOPE requirement to provide PIDs and therefore dynamic and predictable URLs for content and metadata proved to be the final straw. Nevertheless, replacing Fortis raised a lot of concerns. The software had managed their digitization, cataloguing work, and public reading room access, and all these functions would have to be supported in a new system.

The replacement software has been developed in house based on a LAMP stack (Linux, Apache, MySQL, php) and has managed to sustain most of the internal workflows already in place with the added benefit of more

control over databases, metadata schemas, and digital object master and derivative formats. The transfer of metadata and content has also been facilitated, and the system syncs much more easily with their web front end.

FMS's digitization program has not changed. Readers still have access to digitized collections only (with a very few exceptions) and digitization remains a central part of the work on each collection. What has changed are the tools used, the control over processes, and the type and amount of information to which staff now have direct access.
Currently new images and textual material are digitized as uncompressed TIFFs and JPEG access copies and thumbnails are generated. (OCR has not been systematically implemented as they have mostly manuscript material.) Audio is stored in uncompressed WAV files and MP3 derivatives are created. Video varies according to source, but masters are generally DV files and FLV derivatives are created.

Regarding legacy content, FMS still has a few legacy formats from the 90s, mainly PCX and B/W TIFF Group 4. For every image that only has 1 bit depth (B/W), they store the original as a master and generate 8 bit PNGs as derivatives. Files are stored in a conventional file system, RAID hardware and with routine backup to a separate server. For each digital object, they currently generate a unique identifier, MD5 checksum, and store some technical metadata extracted from the exif header (bit depth, length/size, DPI, mime type, date of creation, etc).

Filenames are based on the unique identifier of the document, page order, digital object unique identifier and derivative type.

```
Example file name (derivative 2 of page six of document 05112.003):
05112.003_p0006_id000823493_D2.jpg
```

FMS have already created a PIDs based on Handle, but they still must implement a tool for communicating with the HOPE PID Service to bind PIDs to resolve URLs.

The experience of FMS reveals how an institution deeply committed to its system can still start afresh in response to changing circumstances and needs. Their task was no doubt eased by their in-house technical skills honed through years of adapting and extending their proprietary software. FMS stands out among HOPE partners as a uniquely autonomous organization; in this case, it proved to be their strength.

1.4. Framework(s) of the Social History Institution: References

Bradley, Kevin, Junran Lei, and Chris Blackall. *Towards and Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archive and Preservation System and on Related Software Development.* Paris: UNESCO, 2007.

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System. CCSDS 650.0-B-1 Blue Book.* Washington D.C.: NASA, 2002. (public.ccsds.org/publications/archive/650x0b1.PDF)

Data Asset Framework (DAF). (www.data-audit.eu)

Data Seal of Approval. (www.datasealofapproval.org)

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). (www.repositoryaudit.eu)

Jantz, Ronald, and Michael J. Giarlo. "*Digital Preservation: Architecture and Technology for Trusted Digital Repositories.*" *D-Lib Magazine 11*: 6 (June 2005). (www.dlib.org/dlib/june05/jantz/06jantz.html)

Keller, Paul. "Copyright." In: *Business Model Innovation Cultural Heritage.* Ministry of Education, Culture, and Science: Amsterdam & The Hague, 2010.

Nestor Working Group for Trusted Repositories. *Catalogue of Criteria for Trusted Digital Repositories, Version 1.* Frankfurt Am Main: Network of Expertise in long-lerm STORage, 2006. (www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor)

RLG-NARA Task Force. Trustworthy Repositories Audit & Certification: *Criteria and Checklist, Version 1.0.* Dublin, Ohio: OCLC, February 2007. (www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories)

RLG-OCLC. Trusted Digital Repositories: *Attributes and Responsibilities*. RLG-OCLC Report. Mountain View, Calif.: RLG, May 2002. (www.oclc.org/research/activities/past/rlg/trustedrep)

2. The HOPE Federated Repositories

"Federated Archives are conceptually Consumer-oriented. In addition to the Local Community (i.e. the Designated Community served by the archive), a Global Community (i.e. an extended Designated Community) exists which has interests in the holdings of several OAIS archives and has influenced those archives to provide access to their holding via one or more common finding aids."[5] At present social history collections are accessible only in a disconnected way. Localized, idiosyncratic, and uni-lingual catalogs and finding aids often hamper the discovery experience of a wider audience of social historians as well as the general public. Digital humanities tools fostering innovative research on a wide body of material cannot be effectively used on social history content due to the lack shared standards and practices in the domain. Even when shared standards are adopted, disparate ICT infrastructures continue to work against access and usability—creating information silos at the local level. HOPE provides the opportunity for social history institutions to take part in a cooperative effort: institutions from ten countries have agreed to implement a federated digital repository infrastructure as a short-term goal of the project.[6] To sustain this endeavor, they will: 1) ensure interoperability by providing metadata in major domain standards, harmonizing key values, and assigning globally unique persistent identifiers to managed content; 2) coordinate access management through the data supply, discovery, and delivery process; and 3) embark upon long-term digital assets management. To further pool resources and align practices, several content providers (CPs) are participating in a common PID web resolver service and a long-term storage solution for digital content. The HOPE project is at its essence an attempt to implement the OAIS reference model's federated archives infrastructure with a few key shared functional areas.

The envisaged benefits are:

- Responsiveness to changing community needs: Federated archives have a strong dependence on their user communities. Today's users turn more and more to large-scale discovery services and are less likely to seek out local catalogs to browse and discover content. The ability to disseminate content to selected discovery services based on changing user requirements is key.
- Contextualization of collections: Item-level representations of digital objects on discovery services offer a fragmented view of larger collections; such content risks being overlooked or "orphaned". This can be prevented if institutions with a similar profile act with a common approach. Clustering content from a single sector may also serve to reveal lost or hidden connections.
- Integration of heterogeneous content: Local and proprietary systems do not foster integrated access to collections with a highly multilingual, multi-domain, and transnational profile.

[5] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System. CCSDS 650.0-B-1 Blue Book* (Washington D.C.: NASA, 2002): 6-4.
[6] Heritage of the People's Europe (HOPE), *High-level Design of the HOPE Architecture* (2010).

- Removing data lock-in barriers: A flexible and open architecture enforces integration and makes a strong case for open source solutions at the local social history repositories currently suffering from strong vendor dependence. Storage and dissemination of metadata in widely used standards outside of local systems also unlocks data for other uses.
- Technological innovation: Raising the visibility of social history data requires a unified system of identification, shared vocabularies, common authority records, and a single data model as a common point of reference. Such requirements are supported through innovative technological solutions.
- Sustainability and contingency planning: the federated infrastructure builds a solid foundation for institutions lacking resources. It offers potential savings on storage, data curation, and other service provisions.

2.1. Content Profile

The HOPE social history institutions collect all types of records and publications belonging to transnational social movements, non-profit organisations and global NGOs, national political parties, and private individuals or families–none falling into the category of public records. HOPE CPs define their own Collection Policies, broadly delineating the types of materials they accept and the range of sources. Most do not use quality-based selection criteria on donated physical material or digital content—although in cooperative digitization projects they may dictate the quality of master and derivative files produced. As private institutions they are, simply put, not in the position to dictate the quality of acquired materials. And as they often solicit overlooked and otherwise un-collectable content, they tend to draw a large proportion of ephemera, grey literature, informal works, and perishable material. On the other hand, in keeping with their legacy and mandate, they are highly selective regarding the topical scope of the materials they accession, keeping their institutional and domain profile well distinguished from state institutions. As they often treat transnational or pan-European themes, the HOPE social history institutions often hold "non-native language" or multi-lingual collections. Moreover, materials collected—including publications, personal papers, organizational records, grey literature, paraphernalia, films, and visual materials—tend to cross traditional information domains. Collection policies are realized in statutory agreements and deeds of gift laying out the specific acceptance criteria for each collection and stipulating access and reuse conditions. Archival legislation in each country partly covers legal requirements on the records of these private entities. The Freedom of Information Act is only applicable on their collections defined as public records. For the most part, access to their collections is circumscribed by donor requirements, copyrights, and data protection requirements.

The HOPE Content Policy Framework provides guidelines on the provision of digital objects and descriptive metadata to the HOPE System. The main purpose of the framework is to form clearly delineated data sets—by convention HOPE defines them as "collections"[7] —and to create collection-based templates in order to provide data to the

[7] Heritage of the People's Europe (HOPE), "Defining Your Collections," *The HOPE Manual* (2011).

access service, the HOPE Aggregator. HOPE collections are selected based on two criteria: they should be a "set" of digital objects, based on their production or provenance; and they should relate to the broad social history themes established by the HOPE partners.[8] CPs must tag their HOPE collections and respective items with HOPE Themes[9] in order to bring the HOPE Social History Resource together as a coherent body of material, regardless of local classification structure and language. HOPE Themes will cluster the HOPE social history content in Europeana, highlighting the domain and enhancing the discovery-to-delivery experience. The current content policy covers only digital material. This includes: digitized copies of currently held analog material; copies existing only in digital form, as analog originals are out of reach or have been destroyed; and digitally-born, recently created materials.

HOPE assumes that CPs will take the full responsibility for clearing copyrights on the digital objects and metadata that they offer to the federated HOPE service. The HOPE IPR Best Practice Guidelines[10] state very straightforwardly that "all content provided to HOPE MUST HAVE clear copyright policies. In order to guarantee that all resources are available free of possible 'rights' problems (to the end-user), all content providers must explicitly state the license under which they are providing the content." In order to help CPs achieve this, HOPE has established best practices on the procedure itself ensuring that the collection either falls into public domain or is appropriately licensed to and by the content provider. HOPE further requires that CPs assign use rights metadata to submitted content using the Europeana rights values. In this way, HOPE gives CPs a comprehensive understanding of IPR, contributes to the standardization of in-house procedures for identifying rights issues and assigning appropriate metadata, and helps foster a uniform end user experience. Importantly, such an approach not only supports the HOPE service, it also fosters broader content sharing among social history institutions.

Also important, the IPR conditions stated above do not prevent CPs from storing restricted content in HOPE's common safe storage system. As a basic principle HOPE, promotes open access to social history collections and discourages CPs from using technical means to prohibit access and reuse. On the other hand, it acknowledges that these institutions are often prohibited from fully opening up collections, due to copyright, data protection, or donor embargoes. OAIS repositories must support short-term restrictions with long-term access as the fundamental goal. Currently, the HOPE federated service depends on CPs' local repository infrastructure to regulate access. HOPE has developed an access matrix on the reuse of content to provide guidance on local access management. HOPE also recommends that CPs manage access as granularly as possible to support the HOPE Content Provision Framework.

[8] The concept of social history is itself open to interpretation, and is often defined only by its "oppositional" character. It is declared to be concerned with "real life" rather than abstract theories, with "ordinary people" rather than privileged elites, with "everyday things" rather than political events.
[9] Heritage of the People's Europe (HOPE), "HOPE Theme," *The HOPE Glossary* (2011).
[10] Heritage of the People's Europe (HOPE), *IPR Best Practice Guidelines* (2012).

2.2. Designated Community

Central to the OAIS concept is the designated user community. This community must be clearly defined before the submission of content, and the content as represented should be understandable to the community. HOPE has defined a global designated community made up of the following user communities:

- *Social history researchers and curators* currently using the institutional web sites and online services. This is, simply put, the aggregation of the designated communities of the individual HOPE CPs and forms an inner circle of highly specialized professional researchers already attached to the social history institutions and activities through their routine services and focused research projects. They also have the most interest in long-term access.
- *Professional researchers* who visit the IALHI web site to get informed and browse social history repositories. These are researchers connected to HOPE through its umbrella institution IALHI and are composed of IALHI members and their local communities. Long-term access is also important to this group.
- *Informed European citizens* using Europeana for discovery purposes. This community of citizens interested in European cultural heritage has been identified and is currently served by Europeana.
- *Global users* already using of social platforms like Flickr, Youtube etc.

This has been the biggest challenge facing the social history domain over recent years. The specialist users of the past had developed skills to navigate the archival finding aid structure and library catalogs and had familiarized themselves with domain terminology. Many users nowadays are unable to interpret data through traditional structures and representational forms. They struggle when searching in hierarchical or complex finding aids and rarely take advantage of enhanced search functionalities. In order to target specialist users while also serving broader community needs, HOPE has opted to populate various discovery services, ranging from network sites (e.g. IALHI Portal), to broad-based portals (e.g. Europeana), to globally-known social sites.

Sites will be populated with HOPE metadata, including non-copyrighted descriptive metadata and previews, using various internet protocols. Users locating metadata and corresponding previews will be linked through discovery services back to local sites providing richer content description and contextualization as well as the digital object itself, which will be stored and managed through the HOPE compliant repository. Once a user identifies an item of interest, the digital object may be downloaded and used in the end user's own processing environment, where it may be annotated, edited, mined, used for mashups, etc. The HOPE federated repositories do not offer services to individual users beyond the discovery path itself. The discovery path must be supported by unambiguous persisted references to HOPE collections and digital objects on local sites. In their role as curators, CPs must guarantee authenticity and integrity on the digital content provided online.

HOPE likewise guides CPs in adapting their content provision to various discovery services. In their role as curators, HOPE allows CPs to make conscious decisions about which content is pushed/pulled to each service. HOPE's Dissemination Profiles[11] express in machine-readable form each institution's policies to target specific content to a target user group through a particular service or set of services.

2.3. The Federated Model

OAIS compliant repositories focus on both access and preservation; they serve their designated communities by providing access to content in the short and long term. When they federate, OAIS repositories improve access in the short term by improving interoperability and introducing external components, such as shared finding aids that accept descriptive information in a uniform package from each repository. (In the HOPE federated model, the Aggregator does not serve as a finding aid itself but delivers descriptive information and sometimes Dissemination Information Packages (DIPs) to services already used by the designated community.)

The following are key characteristics of OAIS federated repositories:

- *Central Site/Node*: This site independently manages a set of descriptions from many repositories and serves as a union catalogue, offering a combined view of the holdings. Users are sent from the central access site to local repository sites to retrieve digital objects. Such a system is best supported with a standard set of protocols.
- *Unique Identifiers*: In the federation, a repository is responsible for assigning each Archival Information Package (AIP) an identifier which is unique within the larger system. Ideally, such an ID can store location information to direct consumers to the source repository and/or digital object. The use of web-actionable PIDs not only serves this purpose, but also fosters long-term access.
- *User Authentication and Access Management*: If the federated repositories have the policy of restricting access to some AIPs, there is a need for identifying authorized users making requests through the central access site. Likewise, if access-based services are extended to the general public, for example charging for some content, user authentication should be introduced.
- *Additional Shared Functionality*: Repositories may choose to integrate further to share expensive resources: file management for storage, peripheral devices for ingest, or back-up data storage for disaster recovery.

Partly due to the success of Open Access repositories, which handle research publications and relevant metadata, aggregated digital libraries have become ubiquitous. With a strong emphasis on access, aggregated digital libraries aim to improve the visibility and immediate impact of the research products of their communities. Digital library aggregation solutions have two main functions: 1) harvesting records based on a common schema from local repositories, generally using OAI-PMH; and 2) making data

[11] Heritage of the People's Europe (HOPE), "Developing Dissemination Profiles," *The HOPE Manual* (2011).

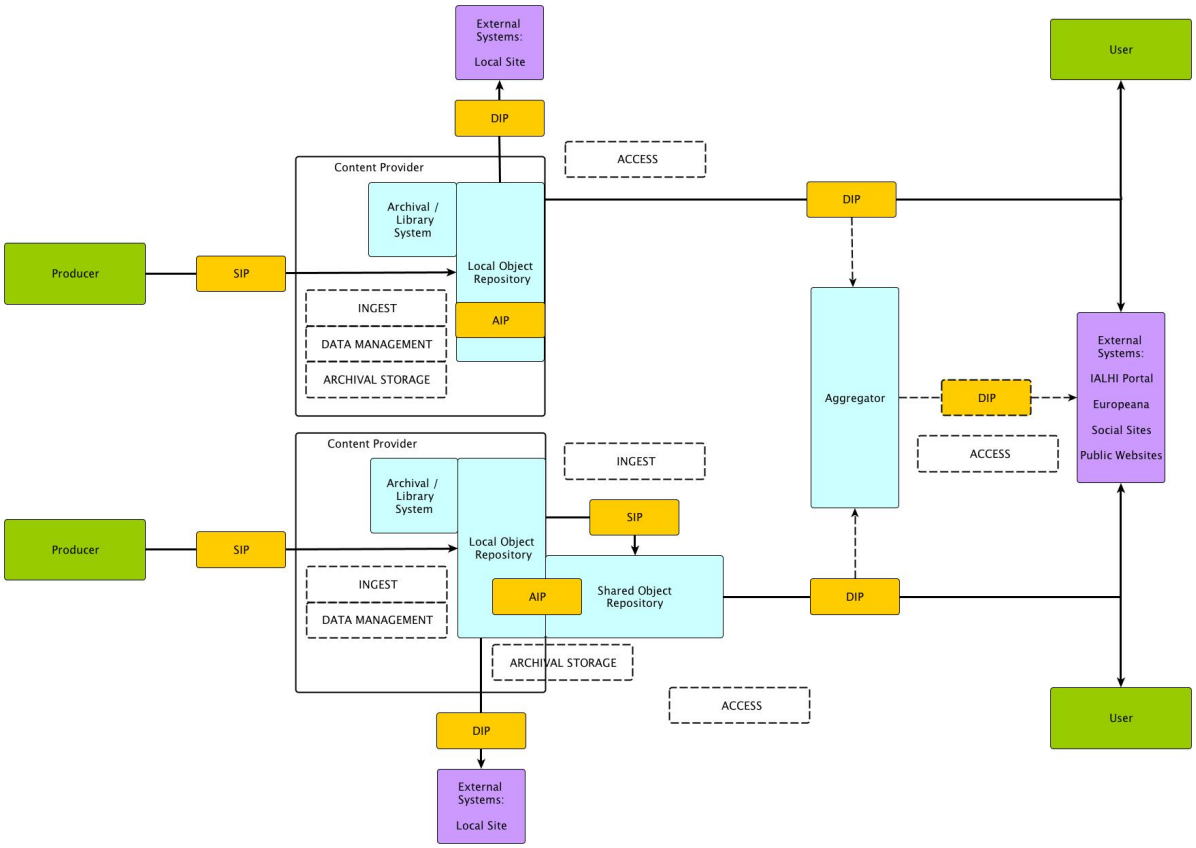
available through one or more common finding aids—either as part of the same service, such as DRIVER, or by pushing data to external services. The aggregated digital library model is in some ways very similar in structure to the OAIS federated repositories model; they both offer a common environment for managing data sets coming from various local repositories. They both focus on shared finding aids. Yet aggregated digital libraries put less emphasis on data integrity, fixity, and validity over the long term.

In a sense, HOPE straddles the two. While built on platforms and protocols developed and used by digital libraries, HOPE's content profile of rare cultural heritage materials and its long-term commitment to its local designated community make preservation an imperative. And while preservation remains out of the scope of the present three-year project, it is never out of sight. If not a driving force, the OAIS reference model at least remains a powerful check on HOPE's activities.

The HOPE federated repositories architecture consists of five core components:

1. *HOPE Compliant Local Object Repositories (LORs)*: Local repositories fulfilling a minimum set of requirements (see below) that are loosely integrated into a general architecture.
2. *HOPE PID Service*: The PID web service helps satisfy the requirement for assigning unique identifiers to each AIP within the federated system. PIDs have likewise become a community recognized standard for supporting long-term persistence of web-based resources.
3. *HOPE Aggregator*: A component of OAIS central access, the Aggregator is an access node which accepts descriptive information in a standard package and makes it available to discovery services. HOPE has split central access into two components, allowing descriptive data sets to be collected in one system and distributed to multiple already existing finding aids.
4. *IALHI Portal and Europeana*: A second component of OAIS central access, the IALHI portal is a primary finding aid or central access site, since it is the main portal for the primary designated community. Descriptive data has been optimized for use on the IALHI portal, which will serve as a union catalogue for digital and non-digital social history collections. Europeana is currently considered a secondary finding aid. However, if Europeana develops the structures and functionality to present the HOPE Social History Resource as a coherent and complex corpus, it may also be considered a primary finding aid.
5. *HOPE Shared Object Repository (SOR)*: An optional shared functionality, HOPE's safe storage facility allows CPs to store master files and perform digital object transformations.

As can be seen, HOPE's underlying architecture is modular, loosely coupled, and extensible, allowing for great flexibility in the management of data and content. While the architecture has clear benefits for HOPE's heterogeneous and growing network of local CPs, it sits uncomfortably with the strict data curation doctrine of the OAIS model. OAIS envisions transparent workflows through coherent functional entities: ingest, archival storage, data management, and access, inter alia. It fails to treat in any detail the problem of synchronization across system components handling a single function.



2-A. Diagram – HOPE architecture & OAIS functional entities

Diagram 2-A shows how the OAIS functional entities are distributed among components in the HOPE System.

- LORs (using a variety of tools and applications):
    - *pre-ingest/ingest*: assign unique identifiers to submitted objects; validate the integrity of objects and check for viruses; ensure presence of long-term metadata; receive a Submission Information Package (SIP) containing descriptive and administrative metadata as well as digital objects; essentially all LORs are responsible for receiving one or more SIPs and packaging them into AIPs and coordinating updates to archival storage and data management entities; some CPs delegate responsibilities to the SOR.
    - *archival storage*: store and manage available Preservation Description Information and in some cases Content Information (in the latter case, error checking, storage management, back up, and disaster recovery may also be performed). Those CPs using the SOR manage Content Information with some overlaps with the SOR.
    - *data management*: manage Descriptive Information and ensure referential integrity across system components.
    - *access*: regulate access to AIPs; deliver DIPs on request.

- Shared Object Repository:
    - *pre-ingest/ingest*: (if not already done) assign unique identifiers to submitted objects; validate the integrity of objects and check for viruses; receive SIPs containing Content Information (including the Digital Object itself and Representation Information); perform transformations; package Content Information for AIPs.
    - *archival storage*: store and manage Content Information, performing error checks, back-up, and disaster recovery. AIPs are ultimately stored and managed by LORs.
    - *access*: deliver DIPs on request.
- Aggregator:
    - *access*: potentially, deliver DIPs on request.

Such a complex architecture must rest on well-defined data flows through the whole system to avoid redundancy or discrepancy. In HOPE, this is supported through the heavy use of PIDs to facilitate the exchange of data and through reliance on common data supply protocols—discussed in more detail in later sections. Yet, LORs remain the linchpin in the system. Flexible and robust local systems are essential and should be extended with tools and applications to support synchronization with the other components. Curation is ultimately the responsibility of local repositories.

2.4. The Federated Model: HOPE Compliant Local Object Repositories

HOPE Compliant Local Object Repositories (LORs) may serve a range of functions, supporting the secure long-term storage and management of content and metadata as well as their regulated dissemination and delivery to local and global designated communities. As the centerpiece in HOPE's discovery-to-delivery path, repositories joining the HOPE federation *must* fulfill a minimum set of requirements. Each LOR:

- must be a networked system connected to the internet;
- must provide online access to digital collections;
- must implement a standard harvesting protocol;
- must be able to assign digital objects and metadata a persistent identifier (PID) or a unique local identifier;[12]
- must make available online at least one derivative file type of the digital object;
- must ensure that the file formats of the digital object can be rendered by widely-supported rendering software (e.g. Web browser, Acrobat Reader, etc.) or else the repository must provide access to appropriate rendering software (e.g. METS-viewer);

[12] Heritage of the People's Europe (HOPE), "Local Identifier," *The HOPE Glossary* (2011).

- must manage access conditions in case the access to the digital object is restricted;
- must record minimum administrative and/or technical metadata to manage the digital objects, such as file format, checksum, statistics.

In the HOPE model, CPs must manage their metadata and digital assets locally. HOPE remains platform agnostic with regard to LORs. Still, proprietary solutions are advised against since the HOPE infrastructure is designed for flexibility, responsiveness, and interoperability. (Service dependencies and data lock-in often prove impediments on the discovery-to-delivery path.) HOPE recommends that LORs be built on open standards and contribute to open source software solutions. The use of dedicated open source object repository platforms, such Fedora and DSpace, is encouraged. CPs using legacy systems based on proprietary software should begin to phase them out. In the meantime, various workarounds may have to be developed to facilitate HOPE compliance. LORs should strive to meet trusted repository criteria as a medium-term goal.

In the HOPE federated model, CPs can choose to extend their LORs with HOPE's SOR storage solution, which is intended to store master files and deliver derivatives on demand. In its present incarnation and with an understanding of the mixed nature of current digital content, the SOR is relatively open without strict quality assurance measures for submitted content. This allows CPs to select a subset of digital objects to store: HOPE objects only or all objects regardless of content scope and access restrictions; simple objects only or also structurally complex objects; high quality master objects, specially-created derivatives, or both. The use of SOR is a "self-governing" process; each CP can define its own business model for using the SOR, evaluating its potential benefits for both short-term and long-term access.

In general, CPs should begin to consider problems of preservation and stable access over the long term. The HOPE architecture offers a range of options for disseminating metadata and delivering content. But to take full advantage of such services, LORs must begin to function as robust preservation and access systems along the lines sketched by the OAIS model.

2.5. The Federated Model: Persistent Identification and the HOPE PID Service

The HOPE federated system must ensure the integrity and good management of submitted content through uniquely identified AIPs. PIDs support duplicate detection, merging, and the reconnection of orphaned records internal to the system. At the same time, PIDs address the problem of broken links as consumers try to locate discovered resources, supporting long-term interaction with digital objects regardless of changes in ownership, location, data format, security, or access protocols. In the context of a federated architecture the consistent implementation and application of PIDs in local repositories can indirectly serve to align in-house workflows, back-end services, and access platforms. The use of PIDs is a clear step toward improved interoperability among social history repositories and greatly facilitates the delivery process over the long term.

As a result HOPE requires persistent identifiers (PIDs) for all submitted descriptive metadata records, entity records (actors, places, subjects), and digital object files.

On the other side, the HOPE Survey revealed that not all HOPE CPs are able to implement a PID system. The most common problem was that local proprietary systems were too expensive to reconfigure for input and storage of PIDs. Other CPs simply lacked the technical expertise to install and maintain a local resolver service. Therefore HOPE developed a web-based service to support CPs in their PID creation, binding, and maintenance work. By implementing a common PID system and a single umbrella resolver service to administer the system, HOPE has endeavored to ease the administrative burden on local CPs, while at the same time providing added functionality through integration with other HOPE components.

The HOPE PID Service is a web-based Handle resolver service with the aim of supporting the management and delivery of HOPE CPs' networked resources. It is an optional service; HOPE CPs may still choose to manage their PID systems locally if they prefer. But for those who choose it, the HOPE PID Service administers the CPs' Handle Naming Authorities through the use of a SOAP protocol for web-based information exchange. LORs may directly communicate with the Service by using SOAP instructions to perform several operations. The HOPE PID Service currently performs so-called "CRUD" operations:

- Creation of PIDs and binds to one or more weighted resolve URLs; to a local identifier;
- Update of PIDs and their bindings;
- Lookup: of bindings via PIDs; of PIDs via resolve url or an attribute (e.g. local identifier);
- Deletion of PIDs.

The service offers Handle system features, such as the choice to use custom naming convention or generate PIDs of (seemingly) random character sequences; and the binding of PIDs to multiple locations (i.e. URLs) and other metadata, specifically CP local identifiers, which can be used for "lookups". As a web-based service, the HOPE PID Service can be integrated with most local collection management or repository solutions. And importantly, a CP can choose leave the HOPE PID Service at any time by installing a local Handle resolution server or transferring to another hosted Handle service.

The HOPE PID Service has been integrated into the larger HOPE infrastructure in several ways. First, CPs are given the option to automatically create Handles during the SOR submission process. Second, once a digital object has been ingested into the SOR, the SOR directly instructs the HOPE PID Service to update its PID binding to an SOR-hosted resolve URL. The resolve URL of any derivative created by the SOR is automatically assigned as a secondary URL, or "location attribute", to the primary object PID. Finally, the HOPE PID Service can communicate with the Aggregator, returning an existing PID for queries on a bound local identifier or resolve URL, or alternatively if no PID currently exists, can automatically create a new PID with bindings to these values. Importantly, through these processes, CPs do not have to store PIDs in their local system but can store and disseminated PIDs using only the HOPE System components. This creates a convenient workaround for CPs that lack the capacity to store PIDs. (Though HOPE

emphasizes that creating PIDs without storing them locally is indeed a "workaround" scenario and not by any means a best practice.)

HOPE seeks to raise awareness on the problems of persistence—both technical/administrative and the underlying institutional commitment—when handling digital objects. HOPE has clarified its policy regarding PIDs and their practical use in the system in the Implementation Guide.[13]

2.6. The Federated Model: Transforming and Disseminating Data through the HOPE Aggregator[14]

The HOPE federation ensures interoperability by harmonizing metadata from various domains and linguistic and institutional settings; disseminates data to a range of discovery services based on CP-defined Dissemination Profiles; offers enhanced search and browse functionality on the IALHI portal; and delivers openly accessible content in widely-accepted formats. In this way, HOPE aim to create a seamless discovery-to-delivery experience.

HOPE relies on CPs to curate their descriptive metadata locally: maintaining clearly defined data sets and collection profiles, ensuring the referential integrity of descriptive entities, providing required metadata elements, upholding the semantic quality of metadata, and following agreed upon character and metadata encoding standards.[15] Such preparatory work, based on the Normalization Guidelines, is ongoing but should ideally be completed before the mapping process is begun. HOPE encourages CPs to employ the widely-used domain specific standards that were used as the basis of the the HOPE Schema: EAD[16], MARC21, and LIDO for visual materials. For those who cannot, the XML mapping sheets allow CPs to map locally-defined elements to the central HOPE Schema. Mapping sheets are also available to map local terminologies to HOPE's normalized lists of values. The HOPE Schema was created with the goal of supplying HOPE's selected discovery services with appropriate data (and if possible to optimize sites' added functionalities); it was developed in close compliance with the Europeana Data Model (EDM).[17] The mapping sheets provide the basis for metadata transformation in the HOPE Aggregator.

The D-Net platform, chosen to serve as the HOPE Aggregator, is already widely used among European digital libraries and archives thanks to: 1) its existing suite of tools for indexing, curating, and enhancing harvested content, 2) its flexible supply and dissemination workflows and support for several search and transfer protocols, and 3) its customized service deployment. In the HOPE federated system, the D-Net based Aggregator is tasked with:

---

[13] Heritage of the People's Europe (HOPE), "Choosing a Persistent Identifier Solution," *The HOPE Manual* (2011).
[14] Heritage of the People's Europe (HOPE), "HOPE Aggregator," *The HOPE Glossary* (2011).
[15] Heritage of the People's Europe (HOPE), "Preparing Your Metadata," *The HOPE Manual* (2011).
[16] HOPE adopts the APENET EAD implementation as best practice. Archive Portal Europe (APEnet), *Mapping towards and normalisation in APEnet EAD: Best Practice Guide* (2011).
[17] Europeana, *Definition of the Europeana Data Model elements, Version 5.2.3* (2012).

- transforming submitted metadata to the HOPE schema;
- normalizing and enhancing descriptive content;
- creating common indexes and authority lists for use on the IALHI portal;
- packaging information to meet the requirements of various discovery services;
- and disseminating descriptive metadata according the specifications laid out in Dissemination Profiles.

The Tagging Tool, especially developed by the D-Net Service for the HOPE project, is one of the features of the HOPE System; CPs can use the tool through the Aggregator's administrative interface to assign a common list of HOPE social history themes at both the collection and item levels. The themes will be indexed, searched, and presented alongside submitted metadata and will serve to further unify the heterogeneous and multilingual HOPE collections. The HOPE authority files have not yet been implemented, but the benefits of such added functionality would be similar. Importantly, the Aggregator makes it possible for CPs to reimport transformed XML files, which means that enhanced descriptive records and authority files could be reintegrated into local systems. Still, it is important to remember that the HOPE Aggregator is first and foremost an access node and should not be used in place of local collection management systems. Metadata curation is the purview of CPs, and unnecessary service dependencies could threaten future preservation efforts.

The Aggregator also offers a customized Export Service capable of exporting descriptive metadata and digital content to social sites, such as YouTube, Flickr, and Scribd, in coordination with the SOR. The SOR has been configured to produce content to specification for HOPE's target discovery services and works together with the Aggregator to push content on demand. Thus, at present this feature can only be used by CPs using the SOR. Here again, it is necessary to underscore the importance of PIDs as the glue which binds the HOPE federated repositories. In the present context, PIDs not only serve to uniquely identify HOPE resources but also allow system components to locate and transfer these resources to external systems.

It should be noted that the services performed by the HOPE Aggregator are by no means static. While the HOPE federation's current requirements are well on their way to being met, the HOPE data model, its attendant schema, and required services will continue to evolve with the emergence of new standards and technologies and in keeping with the changing content profile and designated community. It will be necessary for the HOPE federation to make provisions for such long-term administration of its standards and policies.

2.7. The Federated Model: Secure Storage in the HOPE Shared Object Repository

The HOPE federated repositories aim to store and make available digital master objects and derivatives. Ideally, HOPE LORs can also ensure the fixity and integrity of objects in their care. In its current form, the HOPE Shared Object Repository (SOR) is designated to serve LORs as a secure storage module for digital master files and to transform masters

into derivatives for delivery to external services. In the future, the SOR should become a proper Archival Storage facility which secures the entire set of digital objects and administrative metadata making up the AIP (including Representation Information and Preservation Description Information) and supports emulation and transformation. Descriptive metadata will remain under the direct management of LORs.

HOPE CPs have opted to develop and deploy the SOR primarily in order to pool the cost of technical development and to share expensive storage hardware and back-up services. At present, the SOR has introduced: 1) a scalable safe storage solution with remote back up and 2) a transformation module. Additional features include: dynamic interaction with other components of the architecture through web APIs; automatic creation and administration of PID bindings for submitted objects and derivatives; a function to enable the management of compound objects; a potential payment module; and the possibility to function as a stand-alone "light weight" LOR in full synchronizing with the master SOR.[18]

The use of the SOR in its several capacities will require adjustments in local workflows which may also affect the internal architecture of LORs. The SOR provides an administrative interface which enables CPs to control interactions with both the SOR and HOPE PID Service. Through the administrative interface CPs may set up one or more user accounts to access the SOR Staging Area, a pre-ingest space for storing and preparing files for ingest. Digital objects may be uploaded to the Staging Area via FTPS and arranged in a folder structure by HOPE Collection. In order to ingest objects into the SOR, a so-called SOR XML Processing Instruction is required containing instructions and discrete data sets with a limited number of required administrative metadata elements for each item; the Processing Instruction may be created by the CP itself or automatically generated by the SOR through the administrative interface.

For the digital objects in the SOR to be uniquely identified and accessible by HOPE System components, each submitted object needs to have a PID assigned and provided on the Processing Instruction. Following ingest, the SOR communicates with the HOPE PID Service to update bindings of submitted (or deleted) material; the SOR also provides alternative bindings to available derivatives. In this way, the resolve URL of any derivative created by the SOR is automatically assigned as a secondary URL, or "location attribute", to the primary object Handle, as described below:

```
<ns2:pid>10622.1/EU:ARCHIVE83:ITEM23:FILE3</ns2:pid>
<ns2:locAtt>
        <ns2:location href="http://www.archivalius.org?id=original83.23.3"
                view="master"/>
        <ns2:location href="http://www.archivalius.org?id=image83.23.3.jpg"
                view="thumbnail"/>
</ns2:locAtt>
```

[18] Heritage of the People's Europe (HOPE). *High-level Design of the HOPE Architecture* (2010).

In this case, the URL form of the PID for the master file would be written:

```
http://hdl.handle.net/10622.1/EU:ARCHIVE83:ITEM23:FILE3?locatt=view:master
```

while the version for the thumbnail would be:

```
http://hdl.handle.net/10622.1/EU:ARCHIVE83:ITEM23:FILE3?locatt=view:thumbnail
```

and when used without any location attribute:

```
http://hdl.handle.net/10622.1/EU:ARCHIVE83:ITEM23:FILE3
```

it would point to a jump off page presenting all available versions of the file. This system makes it possible for a CP to locate all derivatives created by the SOR based only on the PID of the submitted file. If a CP is unable to implement PIDs, the SOR system will communicate with the HOPE PID Service to assign a PID based on a local identifier; these will be returned following ingest and transformation along with other updated data on the XML Processing Instruction. Updated Processing Instructions can be downloaded by CPs for use in their local repositories. In all cases, synchronization between the SOR and LORs should be carefully monitored.

Though HOPE considered building the SOR on existing repository solutions, such as DSpace and Fedora, the problem of scalability led the technical team to opt for a more modular architecture based on several discrete components; this also facilitated integration with other web-based services such as the HOPE Aggregator. Each component is based on open source applications which are simple, flexible, and configurable. Three web APIs support interactions with the SOR: submission, dissemination, and administration. Modules include:

- *Identification, Authentication, and Authorization System*, applies access restrictions on categories of users and types (and uses) of digital objects.
- *Ingest Platform*, validates submission requests from the submission API. Fixity values (based on the MD5 algorithm) for all ingested files are generated at this stage. Currently, beyond the initial fixity check, the SOR does not perform random error detection or integrity checks, though this should soon be introduced.
- *Digital Object Depot*, stores digital masters.
- *Convert Platform*, handles a wide variety of formats and creates derivatives in most current web-standards.
- *Derivative Storage*, manages the derivatives created by the Convert Platform. (In fact, digital masters and derivatives are stored in the same storage mechanism, but the Derivative Storage interacts with the Cluster Manager. This multi-node setup of the storage ensures higher capacity and direct connection between Delivery Platform and Convert Platform and helps prevent bottlenecks in retrieving derivatives from the Object Depot.)

Essentially, the SOR performs several of the functions of the OAIS Ingest and Archival Storage entities: receiving submissions, performing quality assurance and virus checks,

converting files and packaging content with metadata, and managing the storage hierarchy. Disaster recovery is likewise ensured through the MongoDB replicaset, which replicates data and stores it at a remote location with an external service provider. CPs can set their own replication and storage policies through the administrative interface. The SOR handles in effect what Hitchcock, Brody, et al. refer to as the "base preservation package" (or bitstream preservation). In preservation terms, this is the first line of defense and, perhaps more relevant for the future of HOPE, often the first phase in the development of a full preservation suite.[19]

In its current state, the SOR does not employ file format identification, validation, extraction tools to collect and confirm embedded data such as would ensure appropriate rendering of objects. On the other hand, the SOR will soon support compound (multi-file) objects through the ingest and storage of METS files containing structural metadata along with the content. Quite understandably, the SOR lacks long-term preservation services—which lie outside the scope of the current project. There is as yet no replace media function to migrate masters (i.e. transform AIPs); at present CPs would have to migrate masters in house and resubmit content under the same PID. Such a function would depend on a rich set of technical and provenance metadata which the SOR still lacks the capacity to store. In essence, the SOR currently supports only part of the Representation and Preservation Description Information. "While replication and storage can provide some support for preservation, it is not a complete solution in the longer term because the effects of format obsolescence require more expert support."[20] Such preservation services should be a longer term goal of the SOR.

Finally, the SOR generates derivatives for access purposes and makes them directly available through the jump-off page. Derivative types are scaled according to the needs of the delivery platforms and services, and access is regulated to prevent the possibility of IPR infringement. For example, Europeana requires a standard preview in a specific size which corresponds to Derivative 3 in the HOPE Access Matrix. Meanwhile Derivative 2 is a standard DIP package, a low-resolution representation for online access and reuse. Derivative 1 is a high-quality copy strictly for permission based reproduction, and as such most commonly required for visual materials. Each derivative type is connected to a range of preferred formats produced by the Convert Platform, allowing a certain freedom of choice for each CP. Through the use of PIDs, content may be transferred to local sites, the Aggregator, or external services for the creation of customized DIPs. Alternatively, users may be sent directly to the SOR to view content and delivery options.

The SOR is built over a flexible architecture that is intended to scale and evolve with the needs of the HOPE federation. In its current incarnation it supports the core services necessary to meet the short-term requirements of the HOPE project, ensuring that objects are secure and available in an appropriate form for delivery to external services and users. The requirements for the SOR will surely change as content and dissemination profiles evolve, supported formats become obsolete, and the HOPE Best Practice Network

[19] The authors suggest that the SOR is following a common course. "Given the low age of most IRs, [format obsolescence] has not yet become a major issue, and there are few examples currently of preservation services that go beyond simple storage." Steve Hitchcock, Tim Brody, Jessie M. N. Hey, and Leslie Carr, "Digital Preservation Service Provider Models for Institutional Repositories: Towards Distributed Services," *D-Lib Magazine* 13: 5 / 6 (May/June 2007).
[20] Ibid.

(BPN) turns its attention to issues of preservation over the long term. As with the Aggregator, the HOPE federation will surely need to shepherd the development of the SOR.

2.8. Sustainability of the HOPE Federated Repositories

"[T]he autonomy dimension is a key one for interacting archives, determining the ease with which each can effect changes in the nature of the association and the impact/penalty to each recovering full autonomy."[21] Whether because of their cultural, economic, or organizational context, social history institutions can have difficulties meeting common policy requirements. Their autonomy remains paramount. OAIS federated repositories can frame membership in such a way that member repositories can leave without notice or impact. In such a "free association", membership would rest on technical and administrative compliance alone, and failure to comply would mean that a member was opting to leave the association. On the other hand, such an association works against the establishment of common the standards and protocols which underpin the seamless discovery to delivery experience.

During the first two years of the project, it has become clear that the permanence of the HOPE federation cannot be guaranteed unless the association is re-negotiated and the autonomy of CPs is restrained. The HOPE Consortium is the group of content providers in the HOPE project who have made contractual commitments related to the governance structure of the project, the rights and obligations of participants, licenses on the results of the project, and dispute resolution within the federation. The agreement will terminate when the EU funded project ends. As a result IALHI, a loose association of social history institutions that has functioned for more than 40 years, decided to establish the IALHI Foundation. The foundation aims to set up an organizational structure to foster the long-term objectives of HOPE, including the maintenance and expansion of the HOPE services and Best Practice Network. While content providers are ultimately responsible for the quality, accessibility, and longevity of the HOPE Social History Resource, the foundation can shoulder some of the burden by setting in place policies and procedures in compliance with Trusted Digital Repository audit criteria to ensure stability and trustworthiness in the continued operations. The foundation will be legally entitled to interpret the evolving needs of the target community and content providers into relevant practices and services as well as to convey the HOPE vision and its link to the social history domain. As such, it will serve as a backbone of HOPE's future efforts.

2.9. The HOPE Federated Repositories: References

Archive Portal Europe (APEnet). *Mapping towards and normalisation in APEnet EAD: Best Practice Guide.* 2011. (www.apenet.eu/images/docs/apenet_mapping_normalisation_guide.pdf)

[21] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System. CCSDS 650.0-B-1 Blue Book* (Washington D.C.: NASA, 2002): 6-9.

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System. CCSDS 650.0-B-1 Blue Book*. Washington D.C.: NASA, 2002. (public.ccsds.org/publications/archive/650x0b1.PDF)

Europeana. *Definition of the Europeana Data Model elements, Version 5.2.3.* 2012. (pro.europeana.eu/edm-documentation)

Heritage of the People's Europe (HOPE). *The HOPE Glossary*. 2011. (igwiki.peoplesheritage.eu/index.php/Glossary)

Heritage of the People's Europe (HOPE). *The HOPE Manual*. 2011. (igwiki.peoplesheritage.eu/index.php/The_HOPE_Manual)

Heritage of the People's Europe (HOPE). *High-level Design of the HOPE Architecture*. 2010. (www.peoplesheritage.eu/content/news.htm)

Heritage of the People's Europe (HOPE). IPR Best Practice Guidelines. 2012. (www.peoplesheritage.eu/content/news.htm)

Hitchcock, Steve, Tim Brody, Jessie M. N. Hey, and Leslie Carr. "Digital Preservation Service Provider Models for Institutional Repositories: Towards Distributed Services." *D-Lib Magazine 13*: 5 / 6 (May/June 2007).

Manghi, Paolo, Marko Mikulicic, Leonardo Candela, Donatella Castelli, and Pasquale Pagano. " Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System." *D-Lib Magazine 16*: 3 / 4 (March/April 2010).

3. Managing Objects Through Administrative Metadata

3.1. Administrative Metadata

Administrative metadata provides information to help manage a resource according to locally defined needs, to secure its integrity, and to enable it to be accessed and used by the target community.

The concept of administrative metadata is relatively new to the information community, outside of a few specialized fields. It came into its own starting in 2004, when several major papers were released which attempted to sketch out the digital library terrain. To this end, NISO defined three types of metadata, descriptive, structural, and *administrative* - a typology clearly influenced by METS. According to NISO, "administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it."[22] Around the same time, technical guidelines released by the MINERVA project described administrative metadata as "used for managing the digital object and providing more information about its creation and any constraints governing its use."[23] Both emphasize digital object "creation", "management", and "access" or "use".

MINERVA included: *technical metadata; source metadata; digital provenance metadata;* and *rights management metadata* - aligning closely with METS-though it treated preservation metadata and administrative metadata separately. NISO explicitly listed only *rights management* and *preservation metadata* as subsets of administrative metadata. NARA, in its own best practice publication, suggested that administrative metadata comprised: *technical metadata* and *preservation metadata*. As recently as 2008, Richard Gartner defined administrative metadata as "the information necessary to curate the digital item, which includes (not exclusively): *technical metadata* […], *rights management* […], *digital provenance* […]";[24] as he later makes clear, preservation metadata includes specific elements within all these categories of data.

What we can see is both a general consensus in the field concerning the scope and function of administrative metadata and at the same time a lack of specificity around the concrete groups of elements that comprise it. This is evidenced by the lack of any unified content or structural standards for administrative metadata. Moreover, though the term "administrative"-or other words used to define administrative metadata such as "management" or "curation"-suggests a set of concrete activities that will be supported by this metadata, in none of the guidelines and manuals are these activities made

[22] NISO, *Understanding Metadata* (Bethesda, MD: NISO Press, 2004), 1.
[23] MINERVA, *Technical Guidelines for Digital Cultural Content Creation Programmes, v1.2* (Bath, U.K.: UKOLN, 2008), 23.
[24] Richard Gartner, "Metadata for Digital Libraries: State of the Art and Future Directions," *JISC Technology and Standards Watch* (Bristol, U.K.: JISC, 2008), 5-6.

explicit. NARA aptly sums up that "these [non-descriptive] types of metadata tend to be less standardized and more aligned with local requirements."[25]

What is clear is that the concept, which is actually more of an anti-concept (i.e. whatever is NOT descriptive and structural), has been reinforced by the rise in the use of METS. (The fact that administrative metadata is almost always used in reference to *digital* resources also confirms this.) The METS document structure has a section dedicated to Administrative Metadata. This section "contains the administrative metadata pertaining to the digital object, its components and any original source material from which the digital object is derived"[26] and has four discrete sub-sections:

- *Technical Metadata <techmd>:* information regarding creation, format, and use characteristics of the files which comprise a digital object.
- *Intellectual Property Rights Metdata <rightsmd>*: information about copyright and licensing pertaining to a component of the METS object.
- *Source Metadata <sourcemd>*: information on the source format or media of a component of the METS object such as a digital content file. It is often used for discovery, data administration, or preservation of the digital object.
- *Digital Provenance Metadata <digiprovmd>*: information on any preservation-related actions taken on the various files which comprise a digital object (e.g., those subsequent to the initial digitization of the files such as transformation or migrations, or, in the case of born digital materials, the files' creation). This information can then be used to judge how those processes might have altered or corrupted the object's ability to accurately represent the original item.

All can be expressed according to any number of known metadata standards or locally produced XML schemas. The METS Editorial Board has endorsed several community-based standards and several have been developed in recent years to treat both Technical and Intellectual Property Rights Metadata. However, even the METS editorial board notes, "Administrative metadata is, in many ways, a much less clearcut category of metadata than what is traditionally considered descriptive metadata. While METS does distinguish different types of administrative metadata, it is also possible to include all metadata not considered descriptive into the <amdSec> without distinguishing the types of administrative metadata further."[27] One thing that is clear from the METS subcategories, however, is that not all nondescriptive metadata is administrative-specifically, the metadata that supports general system administration and database management functions, including: authentication and use logs; metadata on the creation, modification, deletion of metadata records (though METS does record this information, it is not treated as administrative metadata); metadata on the batch interchange of records; and access and delivery logs.

In the last few years, PREMIS has emerged as the sole standard and schema to more or less address all the subcategories of Administrative Metadata listed in METS. The PREMIS

[25] U.S. National Archives and Records Administration (NARA), *Technical Guidelines for Digitizing Archival Materials for Electronic Access* (College Park, MD: NARA, 2004), 6.
[26] <METS> *Metadata Encoding and Transmission Standard: Primer and Reference Manual, v1.6* (Washington D.C.: Library of Congress, 2007), 23.
[27] <METS> *Metadata Encoding and Transmission Standard: Primer and Reference Manual, v1.6* (Washington D.C.: Library of Congress, 2007), 77.

model is based on four conceptual entities: *Objects*, *Events*, *Agents*, and *Rights*. In most cases, PREMIS entities align cleanly with the METS subcategories:

- PREMIS Object Entity corresponds to METS Technical Metadata;
- PREMIS Rights Entity corresponds to METS Intellectual Property Rights;
- PREMIS Event Entity corresponds to METS Digital Provenance;
- PREMIS Agent Entity has less clear correlation, though it has been suggested that it may also be included within METS Digital Provenance.

METS Source remains problematic. The METS manual lists only PREMIS as "a current source description standard". Given the stated role of the Source subcategory, however, it is not clear how the METS editorial board would apply PREMIS, and we have not been able to locate model implementations. In most cases, it is in fact the domain-specific descriptive standard which captures the media and other physical description information on the source material, though for born-digital or analog AV formats, it is possible that PREMIS could offer possibilities for more granular and standardized physical description.

What is perhaps more interesting is that PREMIS markets itself as a preservation standard. This raises the question: What is the difference between administrative and preservation metadata? According to Priscilla Caplan, though "preservation functions can vary from one repository to another, they will generally include actions to ensure that digital objects remain viable (i.e., can be read from media) and renderable (i.e., can be displayed, played or otherwise interpreted by application software), as well as to ensure that digital objects in the repository are not inadvertently altered, and that legitimate changes to objects are documented."[28] Preservation metadata, explain Lavoie and Gartner, has become necessary due to the nature of digital objects themselves. Unlike their analog counterparts (though magnetic tapes may be an exception to this), digital objects are *technology dependent*, *easily mutable*, and *deeply bound by digital property rights*. All three qualities are amplified due to the brief "shelf life" of storage media and rapid obsolescence of technology. For this reason, preservation metadata must include the following information:

- *Provenance*: Who has had custody/ownership of the digital object?
- *Authenticity*: Is the digital object what it purports to be?
- *Preservation activity*: What has been done to preserve the digital object?
- *Technical environment*: What is needed to render and use the digital object?
- *Rights management*: What intellectual property rights must be observed?

(From: Lavoie and Gartner. *Technology Watch Report: Preservation Metadata*. 2005.)

In essence, the purpose of preservation metadata is twofold: (1) to establish and secure the fixity, integrity, and authenticity of the digital object; (2) to enable present and future users to access, render, and use the digital object and its intellectual content. In the terms of OAIS reference model, this is the same sets of information that make up the so-called *representation information* and *preservation descriptive information* at the core

[28] Priscilla Caplan, *Understanding PREMIS* (Washington D.C.: Library of Congress, 2009), 4.

of the information packages submitted to, archived in, and disseminated through digital repositories.

In fact, administrative metadata as defined by METS: confronts the same problems; has the same scope; supports the same activities with the same overall purpose as preservation metadata. Both support the ingest, transformation, storage, and securing of information packages. The difference may be more a matter of emphasis, degree, and context than anything else. Administrative metadata is generally collected according to local requirements and depending on the aims of a repository at a given moment. Preservation metadata is generally collected as one of several pillars that support a more comprehensive preservation strategy, which likely entails explicit and documented policies related to: storage management; back up; transformation/migration; disaster planning; rights management; and business succession or contingency planning.

With reference to the HOPE project, this means that for the moment it is necessary to collect only the administrative metadata which supports the HOPE service's specific functions:

- To submit, store, and make available over the medium term digital masters and/or digital derivatives;
- To ensure the fixity and integrity of objects after submission to our system;
- To deliver objects in a form that can be rendered in the online environment and understood by our target users;
- To clarify, record, and implement the access and use rights and restrictions over our content;
- And to this we may add, to store information in a manner that will not preclude later preservation activities.

As the scope of the project develops, administrative metadata can eventually be extended and filled out to support a full range of preservation functions and services.

3.2. Administrative Metadata in HOPE: Current Status

HOPE CPs store at present very little administrative metadata on digital objects. A review of the Content Provider Surveys given at the outset of the project reveals that HOPE CPs have thus far concerned themselves primarily with descriptive metadata. A few store structural metadata, generally with links from metadata records to files located in a directory structure on a file server, or in the case of BDIC, a quite complex system of metadata and object storage united by an Excel-based integration file.

Identifiers in some form are used by all CPs to store and manage local metadata and objects, but only three have implemented globally persistent identification systems. FES library use the German namespace in the national library URN system (urn:nbn:de:) with registration and resolver services hosted through the Deutche Nationalbibliothek for its metadata records. OSA use the Handle Service for records in its DSpace repository, though in order to implement Handles for its HOPE digital objects and metadata, it would have to register and administer a second Naming Authority. IISG use PIDs internally for

descriptive metadata records. The French institutions likewise indicated that they might need to employ ARKs in accordance with Bibliothèque nationale de France. Thus use of PIDs is currently limited, and three institutions initially expressed unwillingness to implement PIDs; three others remained undecided.

Digital object file names, on the other hand, serve an important role as identifiers within the HOPE institutions. File naming conventions are thus generally quite developed. For the overwhelming majority of HOPE CPs, root file names are created on the basis of institutional domains – library or archival standards – in an attempt to map digital objects and files against physical collections. Such solutions can include straightforward references to the physical archival units, inventory numbers or library call numbers, media and format specifications, or repository IDs. A few CPs (VGA, IISG, FMS) use automatically generated "meaningless" names, such as randomly created numeric codes.

Some examples from HOPE CPs:

CP	model or sample	syntax
AMSAB	archive number-item number-date.extension	ABC123-123456-DATE.EXT
Génériques	jpeg/FRGNQ_PM_cote.jpg (books of songs)	INSITUTIONID_TYPEID_PAGECOUNT.EXT
CGIL	unique inventory number of the object (document, image, booklet)	123456.EXT
FES-Archive	sub-collection-signature of the object	COLLECTIONID-123456.EXT
UPIP(BDIC)	BDIC_CD_AUD_0000	INSTITUTIONID_TYPEID_12345.EXT
SSA	SSA Sozarch_F_5011_Na-0001-452	INSTITUTIONID_DEPTID_123456_ MEDIATYPEID_SIZEID_ REPOSITORYCOUNT_ITEMCOUNT

File and directory naming systems depend on several factors: is the object digitally born or was it digitized; how much flexibility is allowed by the cataloging software; are there any IT system or network limitations; are the files being exchanged; what options are allowed by the back-up or storage system? In all cases the goal is the same, to uniquely identify files within the institution. Many institutions rely heavily on file names and directory paths as structural metadata within their digital object management systems—a dubious practice. OAIS suggests that such "packaging information" is by nature transitory and cannot serve as content and preservation description information.[29]

Technical Metadata is scarce. Three institutions store fixity information. SSA store file checksums in its IMS client using MD5 algorithm. The VGA system also has a so-called "M-Box HealthCheck", though they do not specify which method and algorithm is used to

[29] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System, CCSDS 650.0-B-1 Blue Book* (Washington D.C.: NASA, 2002): 4-30.

support this. OSA use checksums only for non-HOPE content stored in a separate repository. Otherwise, three institutions suggest that representation information is stored along with objects. Only FMS provides details, stating that they provide links to external viewers for AV material. SSA and VGA give no specifics in this regard. SSA with their extensive AV collection and dedicated AV repository store the most explicit technical information, including: file size, file date creation, width, height, bits per px. Finally, several CPs store the location of masters on tape or other long-term storage media.

It is notable that, though many tools are available to generate this information, no CP mentions saving information on the file format separately from the file extension. Neither do any explicitly mention capturing technical metadata related to the creation of the digital object. Finally, though checksums are used by a few CPs, digital signatures are not used by any.

Rights Metadata is recorded by about half of the CPs. Though there is no detailed explanation available, it can be assumed that most store information on the copyright owner in a free text field. SSA and OSA also mention storing additional information on restrictions (donor restrictions or data protection), SSA in a separate field. No use of CC licenses is recorded except by FES-Archive for a special project. In final reckoning, most CPs use a combination of technical means (i.e. watermarking, low resolutions copies, or other limits to online access) and basic free text metadata to control access to and use of collections.

Digital Provenance Metadata is not routinely collected by most CPs. Two CPs have implemented repositories which change the file name upon submission. FMS's Westbrook Fortis packages files in a proprietary format and creates a new file name. The name of the submitted file is not saved. VGA's MBox rename files but also store the name of the submitted file. SSA do not mention any alterations in the file name upon submission to their repository. The in-house repositories of all three of these CPs create audit trails of activity. VGA's MBox supports full-scale versioning of both metadata and objects. For these three institutions, only activities which took place external to the repositories, such as the initial creation of digital objects, are not tracked. The range of functionality supported by their repositories thus determines the scope of their digital provenance metadata. Otherwise for the majority of CPs, there is little explicit digital provenance metadata recorded. However, this may be overstated. A combination of informal documentation, such as scan logs, and tacit information on quality control, transformation, and migration policies and procedures likely exists and could be used to create standard metadata when the need arose. Filezilla or other FTP clients may also facilitate event tracking.

More worrying perhaps is the number of CPs that outsource digitization and even repository functions. Eleven of the thirteen CPs note that they outsource digitization at least some of the time. While this is not troubling in and of itself, it does mean that they risk losing the technical and provenance data on the creation of digital masters and derivatives. While information on the software and hardware used to create objects is relatively easy to retrieve even after a lapse of some years' for digitization work undertaken in house, it may be nearly impossible to get from external vendors after the fact. Several institutions likewise note dependence on an external or parent organization for their technical infrastructure. FES archive and library share the services of a central

IT unit. OSA and MSH-Dijon both depend on the universities for such general IT support. TA outsources the preservation of master files. VGA and CGIL both depend on external organizations for their entire repository infrastructure. Though this is not necessarily a disadvantage-often quite the contrary-it may be an additional obstacle to the collection and storage of standardized administrative, and particularly digital provenance, metadata.

3.3. Administrative Metadata in HOPE: Recommendations

As noted above, for the moment it is only necessary to collect the administrative metadata which supports the HOPE System's specific functions, though it is also important that our strategy should not preclude future preservation activities. Given the complex nature of the HOPE System and underlying services, it is best to approach our general recommendations by identifying the function of each HOPE module and ensuring that the administrative metadata necessary to carry out this function is collected.

The **Aggregator** has as its primary function to store and disseminate descriptive information and their related Dissemination Information Packages - in other words to deliver objects in a form that can be rendered in the online environment and understood and used by our target users. For this, it is recommended that the Aggregator maintain sufficient reference, representation, and context information to allow a digital derivative object to be accessed, rendered, and used by our target user group as well as the rights data to support such access and use.

Highly recommended, administrative metadata on:

- Identifiers that are globally unique and resolvable on the web, *description*, *object*, or *file level*;
- File format and size of access derivative, *file level*;
- Copyright, licenses, or other use restrictions, *description or object level*;
- Use, role, or variant of access derivative (e.g. derivative 2, preview, thumbnail), *object level*;

as well as other not strictly administrative metadata that may serve the above purposes:

- Original material type and language (this is generally counted as descriptive metadata but may also serve as representation information), *description level*;
- Granularity of item (e.g. document, periodical issues, set or "file/folder" of documents), *description level*;
- Structural metadata, *object or file level*;
- Access Restrictions, *description or object level*.

Recommended, administrative metadata on:

- Physical characteristics which inhibit access, *description or object level*;
- Access facilitators (e.g. time coding), *object or file level*.

The **SOR** has as its primary function to ingest, store, and make available over the medium term digital master objects as well as to support object transformation (i.e. derivative creation). For this, it is recommended that the SOR store the reference information needed to manage files and objects, the representation information necessary to create derivatives in a format required by portals, and the representation information needed to make objects accessible to target users. It is also recommended that the SOR store fixity information to support routine quality control and provenance information tracking transformations within the system itself.

Highly recommended, administrative metadata on:

- Object identifiers for masters and derivatives that are globally unique, *file and possibly object level;*
- Fixity of masters, *file level*;
- Viewers and players that are not readily available to the average user, specifically AV players (links to external viewers or embed links would also suffice) for masters and derivatives, *file or object level*;
- File format of masters and derivatives, *file level*;
- File size for masters and derivatives, *file level*;
- Use, role, or variant of object (e.g. master, derivative 1, derivative 2, preview, thumbnail), *object level*;
- Audit trail logging transactions with files from the point of submission, *file and object level*;

as well as other not strictly administrative metadata that may serve the above purposes:

- Structural metadata, *file and object level*;
- Access restrictions, *object level*.

Recommended, metadata on:

- Format version and registry information for masters, *file level*;
- Fixity of derivatives, *file level*;
- Submitted master file name, if changed, *file level*.

The **LORs** generally serve a range of functions, which may encompass those above. In the case of non-SOR users, local repositories may support ingest and storage of master files and derivative creation as well as routine quality control. LORs may also support some of the Aggregator functions on their own local sites or may independently export content to portals. For these, they should follow the recommendations above.

Regardless, it is highly recommended that all HOPE LORs *should be able to produce when needed* the reference, representation, and context information (e.g. identifiers, file formats and size, viewers and players, structural metadata, language, type, and granularity) that is necessary to represent objects in an online environment for target users and the rights information that supports access and use. LORs will also play a key role in any future preservation activities. Care should be taken to collect and store relevant technical metadata on the events in the digital life cycle of each object from the moment of its creation. (It is important to note that there are currently no hard

requirements or recommendations on whether and how such information is stored but only that LORs should be able to produce the information if needed.)

3.4. Administrative Metadata: References

Caplan, Priscilla. *Understanding PREMIS.* Washington D.C.: Library of Congress, 2009. (www.loc.gov/standards/premis/understandingpremis.pdf)

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System*. CCSDS 650.0B1 Blue Book. Washington D.C.: NASA, 2002. (public.ccsds.org/publications/archive/650x0b1.PDF)

Gartner, Richard. "Metadata for Digital Libraries: State of the Art and Future Directions." *JISC Technology and Standards Watch*. Bristol, U.K.: JISC, 2008. (www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf)

Lavoie, Brian, and Richard Gartner. *Technology Watch Report: Preservation Metadata*. Oxford: Oxford University Library Services, 2005. (www.dpconline.org/advice/technologywatchreports)

<METS> *Metadata Encoding and Transmission Standard: Primer and Reference Manual, v1.6.* Washington D.C.: Library of Congress, 2007. (www.loc.gov/standards/mets/METS Documentation final 070930 msw.pdf)

MINERVA. *Technical Guidelines for Digital Cultural Content Creation Programmes, v1.2.* Bath, U.K.: UKOLN, 2008. (www.minervaeurope.org/publications/MINERVATechnicalGuidelinesVersion1.2.pdf)

NISO. *Understanding Metadata*. Bethesda, MD: NISO Press, 2004. (www.niso.org/publications/press/UnderstandingMetadata.pdf)

PREMIS. *PREMIS Data Dictionary for Preservation Metadata, v. 2.1*. Washington D.C.: Library of Congress, 2011. (www.loc.gov/standards/premis/v2/premis21.pdf)

U.S. National Archives and Records Administration (NARA). *Technical Guidelines for Digitizing Archival Materials for Electronic Access*. College Park, MD: NARA, 2004. (www.archives.gov/research/arc/techguiderasterjune2004.pdf)

3.5. PREMIS

PREMIS stands for "PREservation Metadata: Implementation Strategies". PREMIS is an international working group established in 2003 to develop metadata for use in digital preservation. The group was "charged to define a set of semantic units that are implementation independent, practically oriented, and likely to be needed by most

preservation repositories". In May 2005, PREMIS released the Data Dictionary for Preservation Metadata; Version 2.0 released in March 2008, Version 2.1 in March 2011.

The PREMIS Data Dictionary defines a core set of semantic units that repositories should know in order to perform their preservation functions. As noted previously, though preservation functions can vary from one repository to another, they will generally include actions to ensure that digital objects remain viable (i.e., can be read from media) and renderable (i.e., can be displayed, played or otherwise interpreted by application software), as well as to ensure that digital objects in the repository are not inadvertently altered, and that legitimate changes to objects are documented.

It is important to note that the PREMIS Data Dictionary is not intended to lay out all possible preservation metadata elements, only those that most repositories will need to know most of the time. It is also important to note that most elements will likely already be present somewhere in a given digital repository and are collected through other repository activities or automatically supplied during the life cycle of the resource; in this sense the PREMIS Data Dictionary serves as a cross-section of existing metadata with a preservation focus. Finally, the PREMIS Data Dictionary is also implementation independent; it defines semanitic units (rather than metadata elements), which may be mapped to any existing schema -though PREMIS XML offers an easy implementation. There is an expectation that when PREMIS is used for exchange it will be represented in XML.

3.5.1. PREMIS: Conformance

PREMIS conformance is relatively liberal, and more important perhaps, than the few requirements for conformance are those things not required. For instance, a repository is not required to support all of the entity types defined in the PREMIS data model. It is also not required to store metadata internally using the names of PREMIS semantic units, or using values that follow PREMIS data constraints. In other words, it does not matter how a repository how a repository "knows" a PREMIS value -by storing it with the same name or different name, by mapping from another value, by pointing to a registry, by inference, by default, or by any other means. So long as the repository can provide a good PREMIS value when required, it conforms. (See Appendix A for discussion of PREMIS and NZLZ.)

HOPE is recommending the use of PREMIS metadata elements within a METS framework for both LORs and the SOR. The PREMIS element set can be extended within the appropriate METS Administrative Metadata sections as needed using PREMIS extension containers.

3.5.2. PREMIS: References

Caplan, Priscilla. *Understanding PREMIS*. Washington D.C.: Library of Congress, 2009. (www.loc.gov/standards/premis/understandingpremis.pdf)

National Library of New Zealand. *Metadata Standards Framework-Preservation Metadata (Revised).* Wellington, N.Z.: NLNZ, 2003. (www.natlib.govt.nz/downloads/metaschemarevised.pdf)

PREMIS. PREMIS *Data Dictionary for Preservation Metadata, v. 2.1*. Washington D.C.: Library of Congress, 2011. (www.loc.gov/standards/premis/v2/premis21.pdf)

3.6. Persistent Identifiers (PIDs)

Persistent Identifiers (PIDs) are unique "names" for entities that are considered have the organizational commitment and technical infrastructure to support them indefinitely. While a PID can be unique in any given context, it is most powerful when it is globally unique in a widely known and used namespace (e.g. ISBN).

With relation to digital content, PIDs are most efficacious when they are sustained by services and protocols which make them actionable or "resolvable" through the internet-binding a resource's permanent identity to its current, but potentially changeable, location on the web and directing requests for the resource itself to the current location. In this case, both PID global uniqueness and the binding with the current location of the resource must be persisted through organizational and technical frameworks. Today there are a several widely-used actionable PID systems supported through various frameworks.

PIDs may serve two key roles in digital repository infrastructures: 1) supporting long-term access to managed digital resources and 2) supporting system integrity by providing stable identification for system entities. Both roles are becoming increasingly important, the latter as individual repositories begin to seek out economies of scale through federated catalogues, one-stop portals, and shared functionality.

Currently, the HOPE System supports any globally unique persistent identifier system that is resolvable through the internet. HOPE has also developed the HOPE PID Service to facilitate the creation, binding, resolution, and management of PIDs for local CPs who choose to use it.

3.6.1. PIDs: Benefits of a PID System

"By ensuring that all references to digital objects are persistent and non-context-sensitive, we ensure that moving resources to different locations will maintain both the references to and from that resource."[30] The two primary benefits is using a PID System are:

- Global uniqueness: context-independent identification supports the referencing of entities through various systems. More concretely, PIDs are independent of any cataloguing or repository software and remain stable when software changes. In aggregator or harvesting services, globally unique identifiers facilitate data supply, helping to identify duplicates and manage updates and deletions.

[30] Sean Reilly and Robert Tupelo-Schneck, "Digital Object Repository Server: A Component of the Digital Object Architecture," *D-Lib Magazine* 16: 1 / 2 (January/February 2010).

- Persistence: 1) Bindings to a resource's current location and other managed information about the resource; along with 2) the technical means to direct requests on the identifier to the location and other bindings, serve as a "key", allowing the identifier to be used in place of the resource itself. More concretely, PIDs can link one resource to another over the long term, preventing "broken links" in the form of altered resources, redirects, or in the worst case "404 not found" messages. "An important strategy to help reduce the danger of failing to retrieve an object is to add a layer of indirection between the browser and the target object... Describing the object to a resolver permits the browser to find a specific instance of it at the last minute." [31]

Both are brought about first through long-term organizational commitment, and second through naming schemes and technological infrastructures which help manage the long-term commitment.

3.6.2. PIDs: Characteristics of a PID System

A robust PID system can have two possible components: a registry and naming scheme that supports the creation of globally unique identifiers; and binding and resolver services that ensure longterm access over the internet to the resources identified by the name. PID systems do not necessarily have both components. When selecting a PID system to support digital repository systems, the following issues should be considered:

- *Identifier actionability*: Is the identifier resolvable on the web? Is the identifier (or can the identifier be expressed as) a URI that can be used directly in web browser or is the mediation of the resolver service necessary?

- *Identifier form and scope:*
  - Is the identifier opaque or semantically meaningful? If it is semantically-laden, are the qualities on which it is based likely to persist? Can the syntax incorporate local identifier systems?
  - Does the identifier syntax support digital object variants and versions? Does it support the relation of component parts? Would it support non-document entities?

- *Supporting Services, Interoperability, Community:*
  - Does the identifier scheme come bundled together with one or more services to create, bind, resolve, and manage PIDs effectively on the internet? Are the services reliable, sustainable, secure, and cost effective? Are the services centralized or locally hosted?
  - Does the identifier system create an technical or administrative dependencies? Are there potential administrative or technical obstacles to using these services?

[31] Tonkin, "Persistent Identifiers: Considering the Options," *Ariadne* 56 (July 2008): 4.

- Does the service support access restrictions for resources not intended for access on the web? How would the service support an identifier for which a resource is no longer available?
- Is the identifier system a formal and well-documented standard? Does it comply with major web standards? Is the system flexible enough to interoperate with or incorporate other schemes? Is it dependent on protocols which may change over time or become obsolete?
- Is the identifier system (with its attendant services) a mature, well-supported, and widely-adopted system with a committed community of users?

HOPE currently recommends that CPs choose a PID system that is actionable through the internet, syntactically flexible to support a variety of uses, and relatively opaque. CPs should choose a widely-used system supported by a robust technical service infrastructure and proven organizational commitment. CPs should opt for a system that has been widely adopted in the cultural heritage community and is inexpensive and straightforward to introduce and maintain in changing financial circumstances.

3.6.3. PIDs: Selecting a System

"There is considerable duplicative effort across disciplines and sectors; although each discipline considers its efforts unique because its underlying data is unique, at an information science level they are often pursuing the same ends by similar means."[32] Three systems currently meet the above HOPE recommendations: PURL, Handle System/DOI, and ARK.

PURL (Persistent Uniform Resource Locators; purl.org): Implemented by the Online Computer Library Center (OCLC) in 1996, PURLs are actionable identifiers in their simplest form; a PURL points to a resolver which returns the current location of a resource in the form of an HTTP redirect. Originally intended to be a transitional phase awaiting URN development, PURLs were designed to be compatible with URN architecture. PURLs may be created using a public PURL server, or they may be created and maintained through a local resolver, which can be implemented using a free software package available from OCLC.

There is no added service cost for implementing and maintaining PURLs and a low technical barrier free of service dependencies, as it is managed locally and built around widely used protocols. However, the technical and organizational infrastructure is also relatively rudimentary. As with all PIDs, once created PURLs are permanent. Bindings to the current location of the resource must be maintained by the PURL creator. If the binding is broken whether by actively deleting the resource or inadvertently changing its location, the PURL and its full history will still be available through the resolver service. In all cases, PURLs respond to queries by returning the appropriate HTTP response codes. PURL supports so-called partial redirection, which facilitates the management of hierarchical resources. Currently, the PURL service does not support access control, and

[32] NISO Identifier Roundtable, March 13-14 2006, "Problem Statement," (Bethesda, MD.: National Library of Medi-cine, 2006).

thus PURLs are primarily useful for openly accessible web-based resources. PURLs are in relative wide use across the library sector, with servers currently hosted by, inter alia, OCLC, the National Library of Australia, the Danish Bibliographical Center, and the U.S. Government Printing Office.

Basic Form:	[Protocol]/[Resolver Address]/[Local Name]
Example:	http://purl.abcd.org/ABC/DEF/200

Handle System / DOI (Digital Object Identifier) (www.handle.net; www.doi.org): The Handle system was developed in 1994 by the Defense Advanced Research Projects Agency (DARPA) and the Corporation for National Research Initiatives (CNRI), which still administers the central site. Handle is a partially bundled service (including protocols, a namespace, and a software implementation) where the creation and maintenance of identifiers and bindings are "outsourced" to a local repository hosting a Handle server. The central Handle service identifies the local server and directs the request to the server for resolution. Like PURLs, Handles are intended to complement URNs, such as DOI, though they can support identifiers of many types. Handle resolver software may be freely downloaded and locally utilized by any institution, though formal participation in the system comes with a small annual fee.

Handle is a moderately inexpensive and robust PID scheme and resolver solution, though administration is more complex than with PURLs. Unlike PURL, the Handle System is not based on HTTP and DNS protocols, though it can function within them. Instead, it maintains its own root server, the Global Handle Registry (GHR) to manage lower-level Naming Authorities. Handles can be resolved through the central resolver service or through HTTP protocols by prepending the hdl.handle.net prefix to the Handle. The system likewise allows granular control by local administrators through a database supporting additional permissions and bindings to multiple locations and other descriptive metadata. Among other things, this allows administrators to specify "multiple resolutions," such as various locations for one resource, (by appending the attribute "?locatt=" along with assigned criteria). This can support nuanced referencing of resource versions and variants. Today Handle provides the technology behind DOI implementation and resolution, allowing DOIs to function as indirect URLs. The Handle System is also used by digital repository software such as DSpace and CNRI's own Digital Object (DO) Repository.

The DOI system was developed in 1997 by the Association of American Publishers and is now managed by the International DOI Foundation (IDF). DOI is a framework supporting structured identification, resolution, and other policies and tools. DOIs can currently be created, bound, and resolved using Handle technologies. DOI Registry Codes are assigned directly through licensed DOI Registration Authorities (RA), rather than by Handle's GHR, and local repositories must work through RAs to create, bind, and maintain their DOIs. Entry and maintenance costs depend solely on the policies of specific RAs but are often quite high.

DOIs are a fully articulated schema built over Handle services, but to take full advantage of DOIs potential requires a serious financial and professional commitment. Like Handles,

DOIs exist independently of DNS and HTTP protocols, but can function within them by prepending the dx.doi.net proxy server or hdl.handle.net server to the DOI. The system is intended as a generic framework for naming entities of all types, though the focus of their model is on intellectual property related parties, resources, and events. DOI aims primarily at semantic interoperability. DOI binds PIDs to INDECs metadata and has developed Handle's support for linking resources into a fully articulated framework for expressing relationships. DOI likewise permits the definition of Application Profiles for specific communities working with similar data formats. DOIs are heavily supported by academic and scientific publishers, national libraries, and international organizations such as the Joint Information Systems Committee (JISC). The system is widely viewed to serve private sector interests. DOI is the first PID system to be officially supported by an ISO standard (ISO 26324:2012, Information and documentation – Digital object identifier system).

Basic Form:	[Handle Naming Authority]/[Local Name]
Example:	hdl:10345/3873
Example URL:	http://hdl.handle.net/10345/3873
Example URL with location attribute:	http://hdl.handle.net/10345/3873?locatt=id:1
Basic DOI Form:	[Directory Code="10"].[Registry code]/[Local Name]
Example:	doi:10.1006/jmbi.1998.2354
Example URL:	http://hdl.handle.net/10.1006/jmbi.1998.2354
Example URL using DOI's proxy server:	http://dx.doi.net/10.1006/jmbi.1998.2354

ARK (Archival Resource Key; www.cdlib.org/inside/diglib/ark/): ARK was created in 2001 by the US National Library of Medicine and is currently maintained by the California Digital Library (CDL). Under ARK, institutions serve as Name Assigning Authorities (NAAN) or sub-authorities. Institutions must assign Name Mapping Authority Hostports (NMAH) where the "mapping" of ARK names to actionable URLs, or resolution, is provided. The purpose is to clearly separate the role of PID creators from that of the service providers that do the name mapping and resolution-though these are often the same institution. An NAAN may assign several NMAHs and a NMAH may serve several NAANs, but on all accounts the NMAH is considered a temporary role. In that sense, a particular ARK PID will only have a single NAAN but may have more than one NMA, even at the same time. The CDL maintains a registry of all NAANs and their currently assigned NMAH service providers. There is no subscription fee or costs; an institution must only contact the CDL to receive a NAAN and simply generate ARKs using software to produce identifiers - open-source solutions are available.

Like PURLs, there is no added service cost for implementing and maintaining ARKs and no service dependencies. Like DOIs and Handles, ARKs theoretically exist independently of HTTP and DNS protocols, but unlike they former, currently they are only actionable

within them. All resolution happens through the assigned NMAH. If an ARK-based URL fails to work because the NMAH is not current, then the current NMAH may be identified through CDL NAAN registry. In ARK, PIDs have fixed multiple resolutions, which can be accessed by appending suffixes, so-called inflections, to the root ARK. These include bindings to the resource location (using the simple PID form), to basic metadata about the resource (using the inflection "?") and to a commitment statement by the NMAH, including change history, future policies, and likelihood of persistence (using "??"). The commitment statement is central to the ARK concept, which rests on organizational commitment as much as identifier syntax or technical infrastructure. ARK's syntax supports optional qualifier strings to differentiate resource variants (using ".") or components (using "/"). ARKs can also be used for restricted resources. ARKs are now used by national repositories such as the Library of Congress, the Bibliotheque Nationale, the British Library, the Library and Archive of Canada, and the National Library of Hungary as well as other organizations such as the Digital Curation Center (DCC), the Internet Archive, and Google, and several major American universities.

Basic Form:	ark:/[NAAN]/[Local Name][Qualifier]
Example:	ark:/13030/tqb3kh8z
Example with inflection to metadata:	ark:/13030/tqb3kh8z?
Example with component qualifier:	ark:/13030/tqb3kh8z/chap3
Example with variant qualifier:	ark:/13030/tqb3kh8z/chap3/fig5.m
Example with alternative variant qualifier:	ark:/13030/tqb3kh8z/chap3/fig5.t
Basic URL Form:	http://[NMAH/]ark:/[NAAN[/[Local Name][Qualifier]
Example URL with given NMAH:	http://bnf.fr/ark:/13030/tqb3kh8z
Example URL with alterative NMAH:	http://loc.gov/ark:/13030/tqb3kh8z

PURLs, Handles, DOIs, and ARKs each have advantages and disadvantages, and their suitability for a particular institution or institutional collection will depend on local factors. While DOI offers a robust schema, it is also requires a substantial professional and financial commitment - likely thanks to its roots in the private sector. ARK and DOI's descriptive metadata requirements would seem a redundancy for institutions specialized in creating descriptive data according to their own domain standards. The possibilities to reflect a resource's structure in the identifiers may also be superfluous for organizations using METS or other structural metadata schema. ARKs require no service commitment and are free to use, but the institutional commitment so central to the ARK concept and the administrative effort underlying this commitment may prove difficult for small organizations. Handles by themselves are less expensive than DOIs and offer a robust service package and, unlike DOIs and ARKs, relatively low barriers to implementation, but the institution must depend indefinitely on a service provider for resolution. PURLs have the lowest financial and technical barriers for implementation and create no dependencies, but they offer little nuance in their syntax and carry no additional metadata-to facilitate access control, for instance.

The HOPE PID Service has opted to use the CNRI Handle System, citing Handle's large user base and strong documentation on APIs, workflows, and service agreements. It was an additional advantage that CNRI is seeking to move the service under the auspices of a large international body such as the UN. (Given that the HOPE PID Service is administered on behalf of several institutions, the steeper learning curve and additional requirements for ARKs may also have proved an obstacle.) HOPE has implemented the Handle system in such a way that each CP using the HOPE PID Service will retain their own Naming Authority, allowing their PIDs and related data to transferred to a local Handle server at any time.

3.6.4. PIDs: Local PID Policies

For local repositories, selecting a system is only half the battle. Putting in place policies and procedures to support the creation, binding, and maintenance of PIDs is more important and less obvious than it might seem. When implementing PIDs in house, institutions should create a PID policy that addresses the following issues:

To what should PIDs be assigned? Current PID systems are designed with flexible syntax and data structures making possible a range of potential applications. An institution must make decisions about whether to assign PIDs:

- To abstract works or to specific manifestations/copies of a work?
- To digital resources only or to physical objects as well?
- To digital masters, to access copies, or to all representations? And what about to individual files or other components?
- To object metadata or solely to objects? Or to some package which encompasses them both?
- To currently maintained versions of resources or to older versions as well?
- To document resources only or to nondocument resources such as the people, groups, places, and concepts that are the topic of linked data efforts (and that have traditionally been managed through authority files and vocabularies)? Or to the rights, events, agents so central to intellectual property and preservation management? Or to the operations performed by the repository itself?
- And finally, to publicly available resources only or to all resources, including those that are internal, private, secure, or restricted?

Here it is important to consider the underlying motivations for applying PIDs. An institution must determine: which resources will themselves be managed and persisted and which are by nature temporary or in flux; which resources are well served by existing local or international identification schemes; which resources will be "networked" in some form and which will remain offline; which resources might be shared, transferred, or exchanged and which will remain local. In the case of archival material or other rare or unique collections, the question of "abstract works" vs. "manifestations" is of course less relevant than it is in the traditional library domain. More relevant may be the relation between the physical originals (what METS refers to as the Source) and digital versions. In OAIS, "identifiers that allow outside systems to refer, unambiguously

to particular content information"[33] are stored as part of preservation description *reference information*. Such identifiers refer to the physical or digital content data objects that are the object of preservation. Thus in OAIS terms, only the base object of preservation must be persistently identified, whether it be an analog original, a master object (potentially composed of multiple files), or each manifestation of the original object.[34] PREMIS, with its focus on digital objects, suggests a more atomistic approach. PREMIS recommends that digital repositories generate and store persistent identifiers on all representations, files, and even bitstreams-depending on the level that the repository stores and manages objects. While use of persistent identifiers is permitted for other PREMIS entities-i.e. intellectual entities (i.e. descriptive metadata), agents, events, and rights-, it is not explicitly recommended.[35]

In short, an institution will have to decide if the "resource" as envisioned is an abstraction that includes a package of files and related metadata (something akin to a METs digital library object or an OAIS Information Package), or, on the other extreme, if each concrete manifestation of a work-be it a variant, version, component, or metadata-should be considered a resource in its own right. The prior would help enforce a more rigorous data model and, if well conceived, could ease administration; the latter is perhaps more concrete and comprehensible and might offer more flexibility to adapt to changing circumstances. In most cases, an institution will fix on a model somewhere in between the two extremes, often depending on current data structures, supporting management systems, and well-worn local identification schemes. Though less pure, these systems are often the most intuitive and feasible to sustain. For archival and other unique material, institutions generally choose to assign PIDs in some combination to: one or more digital representations of an object, each separate master and even derivative file, a descriptive record, and/or the original physical object-for which the descriptive record may serve as a digital proxy. In any case, institutions should seriously consider assigning PIDs to restricted as well as open resources-it is the access restrictions in this case which are temporary, the resources themselves will persist.

What form should PIDs take? Though PID syntax is broadly set by the system that is chosen, each of the systems described above has a [Local Name] element. In the case of Handle/DOI there is also the possibility to bind a single PID to multiple locations and assign attributes, while in ARK, a single PID can resolve to a digital object, descriptive

[33] Notably, there is no requirement that such identifiers be actionable on the web. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System, CCSDS 650.0-B-1 Blue Book* (Washington D.C.: NASA, 2002): 4-28. Both PREMIS and the OCLC/RLG working group distinguish between local and "global" identifiers in this regard. Global identifiers or those which name a class of objects, are persistent and unique but not necessarily unique to a repository object (e.g. ISBN). Thus, both support the use of multiple identification systems for objects as well as, in the case of PREMIS, for agents. OCLC/RLG Working Group on Preservation Metadata, *Preservation Meta-data and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects* (Dublin, OH: OCLC, 2002): 29-33.

[34] Bibliographic or other descriptive information is generally viewed by OAIS as both reference information in its own right and otherwise mutable, depending on the changing needs of the target users. As such, OAIS does not specifically suggest assigning persistent identifiers to descriptive metadata.

[35] Interestingly, though supporting in general the "linked data" philosophy, PREMIS nevertheless favours the storage of literals rather than URIs to reference values from controlled vocabularies. In this case, documentation on the controlled vocabulary should be available through the repository or its supporting organization. Another option PREMIS notes would be to store the values within the system itself and link data through identifiers internal to and under the long-term control of the repository. PREMIS, *PREMIS Data Dictionary for Preservation Metadata, v. 2.1* (Washington D.C.: Li-brary of Congress, 2011): 18.

metadata, or a commitment statement. ARK's syntax also permits the expression of variants and components as qualifiers. An institution must determine:

- What the relationship will be between the PID Local Name element and other locally-supported identifiers, such as accession numbers, ISBN numbers, call numbers, reference codes, and system identifiers;
- Whether to express entity relationships through PIDs, and if so, to do so through adaptions to the Local Name root or through PID system syntax (e.g. qualifiers, attributes, and inflections);
- How/when to use the possible forms for a single PID-more precisely when it is necessary to use the actionable form and when to simply use the root form.

The relationship between PIDs and local identifier systems is tricky. It is common practice to use an established internal identification system as the source of Local Names, and there is some logic in it. Basing the PID local naming convention on an existing system has several benefits: 1) it eases workflow and administration, often saving an additional lookup step; 2) it allows an institution to infer the form of the PID; this can ease the stress on internal workflows-allowing the institution to create the PID in their local system before formally registering it with the PID service (or to forego storing the PID value at all); and 3) it can reflect the institution's broader data model and relation between various entities, without the need for introducing PIDs throughout. Given these arguments, perhaps the better question is not whether existing identifiers should be used, but which. The main criteria is, that the internal identification system should be locally unique and as stable and persistent as the intended PID; once created PIDs cannot be changed. In this case, semantically less meaningful names are generally preferable. If no stable local identification system exists, then the PID service can generate a random Local Name.

The next question is whether to express entity relationships through PIDs or to keep PIDs atomized and semantically opaque. One disadvantage (or advantage, depending on the aim) to using a standard syntax to express relationships is that PIDs can be guessed by end users. "Names designed so that logical changes have logical consequences have been called 'hackable identifiers'."[36] If the institution hopes to control access passively-by not publishing PIDs for restricted or internal resources-, this may cause problems. If an institution is basing PIDs on a local identification system which itself expresses relationships, then this cannot be avoided, though embedding relationships in an idiosyncratic local system, may be less hackable than expressing structure through the universal syntactical rules offered by PID systems.[37] As a rule, if the institutional data model is likely to change over time, it may be better to avoid expressing relationships through PIDs. Otherwise, use of multiple bindings (such as with Handle's attributes) may allow an institution to capture resource relationships without fixing them permanently into the PID. In general, it is recommended to explore the possibilities in the chosen PID system and optimize built-in features to ease administration, while at the same time

[36] J. Kunze and R. P. C. Rodgers, *The ARK Persistent Identifier Scheme*, Internet Draft (2008).
[37] In the case of ARK, the use of qualifiers also presupposes the existence of related resources. Thus, ark:/12025/654/xz/321, actually contains three separate ARKs: ark:/12025/654/xz/321, ark:/12025/654/xz, and ark:/12025/654.

constructing a naming convention that can be sustainable, maintainable, and appropriately opaque.

Example resource PID:	pid:10891/12345abcd
Example related resource PID, with relationship expressed through change to Local Name root:	pid:10891/12345abcd_master_001
Example with relationship expressed through attribute:	hdl:10891/12345abcd_001?locatt=level:master
Example with relationship expressed through component and variant qualifiers:	ark:/10891/12345abcd/001.master

Finally, current thinking suggests that PIDs should be used in the root form whenever possible, basically in all contexts where actionability is not important. As noted by Hilse and Kothe, "The integration of persistent identifiers into URL strings is risky: it introduces problems if those URLs are not clearly marked as containing a certain encoded identifier. In this case, the URL cannot easily be converted at a later point in time because it is hard to determine which element is, in fact, an encoded identifier."[38] This would indicate that the actionable form of a PID should be produced at the latest possible moment in the workflow and only when needed.[39] In any case, whenever the root form is used, it is necessary to reference the namespace, either as part of the syntax, e.g. ark:/13030/tf5p30086k or doi:10.1006/jmbi.1998, or through labels or qualifiers. And of course, a local naming convention should be produced to reflect the above policies, specifying the syntax and permitted characters.

PID policies should also lay out the long-term institutional commitment to PIDs, including contingency planning to counter technological or service dependencies created by the application of a particular PID system. As aptly argued by ARK system creators, in final reckoning "persistence is purely a matter of service, and is neither inherent in an object nor conferred on it by a particular naming syntax. The best an identifier can do is lead users to those services."[40]

In HOPE, PIDs are required for each metadata record and attendant landing page (often the same) as well as each digital object (at minimum a mid-quality access copy is required) submitted to the Aggre-gator. A PID is also required for any submitted authority records on agents, places, events, and con-cepts. For those CPs using the SOR, all master objects supplied to the SOR must have a PID. For CPs that cannot provide PIDs, HOPE provides workarounds based on local identifiers.

There are no syntactical requirements for PIDs submitted to HOPE beyond those imposed by the PID systems themselves. Generally speaking, each resource submitted must have

[38] Hans-Werner Hilse and Jochen Kothe, *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations* (London: Consortium of European Research Libraries, 2006): 45.
[39] Some argue that to support true persistence, the base form should be used in print citations. (See: DOI Handbook)
[40] John A. Kunze, "Towards Electronic Persistence Using ARK Identifiers," *Proceedings of the 3rd ECDL Workshop on Web Archives* (August 2003): 2.

its own PID. Nevertheless, HOPE supports PIDs with multiple resolutions to resource variants; in this case, the PID must be supplied with appropriate attributes. PIDs for landing pages and objects must be submitted to the Aggregator in their actionable form.

The HOPE Aggregator also generates and manages PIDs on its own resources. The Aggregator generates Handles for HOPE agent, place, event, and concept entities, created by culling and enriching terms from submitted metadata records as well as for HOPE theme entities, a unique set of values devised by HOPE CPs. The Aggregator will depend on the HOPE PID Service to administer its PIDs.

3.6.5. PIDs: PID Workflows and Maintenance

"At the end of the day, the only guarantee of the usefulness and persistence of identifier systems is the commitment of the organizations which assign, manage, and resolve identifiers." (Stuart Weibel, Senior Research Scientist, OCLC) Long-term commitments expressed in policies should be supported through formal institutional procedures that answer the following questions:

When should PIDs be assigned? PIDs should be assigned at the earliest possible point after the creation or accession of a resource - it is, in fact, mandatory for the Archival Information Package -, while the (re)binding of PIDs to their networked location can only happen if and when the resource is made available on the internet. In the interval between, PIDs may be registered in the PID service as "unresolved", they can resolve to a stand-in page, or in the most lightweight scenario (and if a predictable naming convention has been observed), PIDs can be assigned locally without recourse to the PID service. In any case, it is recommended that PIDs be based on stable local identifiers which are also assigned early in the creation or accession workflow.

Where should PIDs be stored? The simple answer to this is that PIDs should be stored in local repositories or collection management systems along with the metadata on the resource. PIDs on metadata records can be stored within the metadata record itself while PIDs on files or representations should be stored along with file-level technical metadata and/or in a METs container. However, this simple answer masks the realities faced by many institutions. Proprietary collection management systems often fail to support PIDs. Moreover, many small institutions still rely on file servers, rather than proper digital repositories to store and structure digital objects. In these cases, institutions can create workarounds, such as lookup tables matching local IDs or file names to PIDs, but a good naming convention (e.g. aligning PIDs to file names or to database identifiers or collection reference codes) may also alleviate much of the problem.

How should PIDs be maintained? It is, unfortunately, not sufficient to create and bind PIDs once and then promptly forget them-as many institutions may hope to do. PIDs are a long-term institutional commitment which must be sustained through constant vigilance. It is necessary to put procedures in place to administer PIDs and their bindings over the long term. Procedures should be established for binding and rebinding PIDs to one or more current locations, particularly as part of any server change, repository update, or website redesign. Procedures should also be produced for the removal, update, or replacement of the actual resources identified by PIDs-possibly through the use of place holders, fixed policy statements, or references to alternative resources. With

complex PID systems like Handle, ARK, or DOI, all procedures must take into account the creation and update of all bindings stored with PIDs, in the form of additional metadata, access rights, and service commitments. The policy to use multiple resolve locations and PID qualifiers must also be supported by procedures which manage these through all processes. Finally, the institution needs to draft procedures related to its commitment to the service itself: renewal of subscriptions, software updates, etc.

HOPE does not require CPs to store their PIDs locally, though it highly recommends that they do. For those CPs that cannot, the HOPE System offers workarounds based on local identifiers. The HOPE PID Service requires that CPs register and receive a Handle Naming Authority from the Global Handle Registry, and CPs must manage their own subscription to the service. To facilitate general maintenance of PIDs and their bindings, the HOPE PID Service employs a SOAP protocol.

3.6.6. PIDs in HOPE: Recommendations

As mentioned in Administrative Metadata, the **Aggregator** has as its primary function to store and disseminate descriptive information and objects in a form that can be rendered in an online environment. It is recommended that for all submitted files, objects, or metadata records, the Aggregator should collect/generate, store, and disseminate a PID. Ideally, the PID should be stored in an actionable form that resolves to a HOPE-supported web resource. This web resource may provide content and representation information and/or reference, descriptive, and context information, as applicable. It is also recommended that the Aggregator use PIDs for managing metadata and content internally.

The **SOR** has the role to ingest, store, generate, and make available digital objects. For this, it is recommended that the SOR accept and store PIDs on all submitted master objects and files. The SOR should also generate PIDs for all objects and files it creates through transformation in order to facilitate the transfer of objects between systems. PIDs in the SOR should not be stored in the actionable form.

LORs can serve a range of functions. In all cases, LORs need to be able to produce sufficient reference information to enable users to identify, locate, and interpret objects in an online environment. It is recommended that LORs store PIDs for all digital master objects and files (in the case of compound/complex objects). For SOR users, a PID is required for each submitted master file. HOPE additionally requires actionable PIDs for all metadata records and files submitted to the Aggregator. It is recommended that LORs assign PIDs consistently across types of entities, regardless of access restrictions. Finally, LORs should not store PIDs in an actionable form.

Beyond the recommendation that PIDs be globally unique, relatively opaque, and able to be expressed in a web actionable form, HOPE has no specific requirements or recommendations for the PID system or syntax used. LORs are ultimately responsible for administering PIDs over the long term.

Case Study: Amsab-Instituut voor Sociale Geschiedenis (Amsab-ISG) Implements PIDs

Based in Ghent (Belgium) since its inception in1980, Amsab-Instituut voor Sociale Geschiedenis (Amsab-ISG; Amsab-Institute for Social History) is an officially recognized Flemish cultural heritage institution engaged in archiving documents and other items of progressive social movements and persons. With approximately 50 staff members including a small IT unit, Amsab-ISG manage a collection of over 80,000 library documents, 30,000 image and sound objects, and 40,000 archival records. Since 2004 Amsab-ISG have used the British-Dutch Adlib Information Systems collection management software and currently make materials available through their online catalogue, which is accessible through their website. For Amsab-ISG, the HOPE Project is primarily a means to broaden access to their materials through Europeana and other discovery portals. They also view the Hope Best Practice Network as a potentially useful network that can help guide them over the long term in their object management and preservation efforts. They opted to use the SOR for the latter reason and have participated in its development. They have also opted to use the HOPE PID Service, hoping it would ease the general administrative burden, in particular their synchronization with the SOR.

This being said, the HOPE requirement to supply PIDs for each metadata record and each digital access copy presented an obstacle. The Adlib system, which holds their current archival, library, and visual descriptive metadata, does not provide an easy solution for storing PIDs; Adlib's API for adding additional metadata proved "expensive and flawed". This was one factor in their decision to introduce a new catalogue for their digitized archival and library material. The new system is based on the open source software Collective Access, which offers a flexible data model and support for digital full-text search, both lacking in Adlib. They continue to use Adlib for their non-digital library collections and higher level archival descriptions. Moreover, owing to their commitments in another network, MovE - Musea Oost-Vlaanderen in Evolutie, they continue to manage their digitized visual material through Adlib. It has thus been necessary to create two different PID workflows.

For digitized archival and library materials, Amsab-ISG now automatically import Adlib records into Collective Access, simultaneously breaking down periodical and series records to create item-level records. At the same time, Handles for both metadata records and objects are generated and submitted through a HOPE PID Service SOAP request. (Amsab-ISG have opted not to auto-generate object PIDs as part of submission to the SOR.) For metadata records, their Handle local naming convention is based on their Adlib record identifiers, followed by sequential numbering system for the newly created item-level records-Adlib identifiers were considered to be more robust than the Collective Access numbers, though the use of a system-dependent local identifier might prove problematic if they ever decide to change systems. For objects, their Handle local naming convention is based on the object file names. These can be based on archival reference codes and library call numbers-a potentially more stable local convention than Adlib system identifiers. The HOPE PID Service translates file names in a more conventional form, using CAPS and removing special characters.

Journal title:	Combat : hebdomadaire wallon d'action socialiste
Journal Adlib record number:	400000627
Journal metadata PID:	hdl:10796/A400000627 http://hdl.handle.net/10796/A400000627
Journal year title:	Combat (1961)
Journal year metadata PID:	hdl:10796/A400000627_1 http://hdl.handle.net/10796/A400000627_1
Journal issue title:	Combat (1961)01
Journal issue metadata PID:	hdl:10796/A400000627_45 http://hdl.handle.net/10796/A400000627_45
Journal issue file name:	196101.pdf
Journal issue PDF PID:	hdl:10796/1961_01 http://hdl.handle.net/10796/1961_01?locatt=view:derivative2
Journal issue thumbnail PID:	hdl:10796/1961_01 http://hdl.handle.net/10796/1961_01?locatt=view:derivative3

Their workflow has been eased by their decision to manage only PDF access copies of multipage objects in the SOR-remember, their short-term goal remains access, not preservation.

Thus, they need only store a single object PID for a single item and can easily do so as part of their descriptive metadata. They had not yet considered how their local naming convention for objects would scale to support multipage master files, such as TIFFs, though it would seem that a page suffix (e.g. _001) could easily be appended to the root. For managing PID resolutions, AMSAB-ISG are testing a special MySQL database that registers all newly created PIDs along with their resolve URLs. The system tracks URL changes in Collective Access and sends requests to the HOPE PID Service to update bindings. They are also looking into possibilities for using SOAP requests to update the HOPE PID Service directly.

Visual materials are managed directly in Adlib. This is primarily owing to Amsab-ISG's involvement in MovE, which itself is an Adlib consortium, but also because visual materials were already described at item level in Adlib. For these collections, Amsab-ISG has resorted to a workaround of sorts. Currently, they plan to use an XML batch request to the HOPE PID Service to submit PIDs for visual metadata and objects based on the same naming conventions noted above:

Object name:	Het plan aan de macht
Object Adlib record number:	17960
Object metadata PID:	hdl:10796/A0017960 http://hdl.handle.net/10796/A0017960
Object number:	AF000014
Object file name:	AF000014.jpg
Object file PID: (location attribute assigned by SOR)	hdl:10796/AF000014 http://hdl.handle.net/10796/AF000014?locatt=view:derivative2
Object thumbnail PID: (thumbnail generated and location attribute assigned by SOR)	hdl:10796/AF000014 http://hdl.handle.net/10796/AF000014?locatt=view:derivative3

As they cannot store PIDs directly in Adlib, they are looking into alternatives for storing and managing this information locally. In these cases, it is important to note that Amsab-ISG does not intend to take advantage of the HOPE workaround scenarios, but still intends to track all PIDs locally. At the same time, they acknowledge that their workaround is also not a best practice, and they will continue to seek better solutions.
In their attempts to implement PIDs for their objects and metadata, Amsab-ISG has managed to overcome a number of obstacles. What is perhaps most noteworthy about the Amsab-ISG case is not the complexity of the problem that they confronted, but rather how representative it is of the sector.

3.6.7. PIDs: References

ARK (Archival Resource Key) Identifiers. (confluence.ucop.edu/display/Curation/ARK)

CASPAR: Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. D2301 *Report on OAIS-Access Model.* February 2008. (Doc. Identifier: CASPARRPD230101011_3)

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System. CCSDS 650.0B1 Blue Book.* Washington D.C.: NASA, 2002. (public.ccsds.org/publications/archive/650x0b1.PDF)

DOI System. (www.doi.org/doi_handbook/3_Resolution.html)

Handle System. (www.handle.net/overviews/system_fundamentals.html)

Hilse, HansWerner, and Jochen Kothe. *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations.* London: Consortium of European Research Libraries, 2006. (nbnresolving.de/urn:nbn:de:gbv:7isbn90698450838)

Jantz, Ronald, and Michael J. Giarlo. "Digital Preservation: Architecture and Technology for Trusted Digital Repositories." *DLib Magazine 11*: 6 (June 2005). (www.dlib.org/dlib/june05/jantz/06jantz.html)

Kunze, John A. "Towards Electronic Persistence Using ARK Identifiers." *Proceedings of the 3rd ECDL Workshop on Web Archives.* August 2003. (bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze)

Kunze, J., and R. P. C. Rodgers. *The ARK Persistent Identifier Scheme.* Internet Draft. 2008. (linked from ARK site: confluence.ucop.edu/display/Curation/ARK)

NISO. *Report of the NISO Identifiers Roundtable*. Bethesda, MD.: National Library of Medicine, 2006. (www.niso.org/news/events/niso/past/ID06wkshp/#problem%20statement)

OCLC/RLG Working Group on Preservation Metadata. *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects.* Dublin, OH: OCLC, 2002. (www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf)

Paradigm: Personal Archives Accessible in Digital Media. *Workbook on Digital Private Papers.* 2008. (www.paradigm.ac.uk/workbook/index.html)

PREMIS. *PREMIS Data Dictionary for Preservation Metadata, v. 2.1*. Washington D.C.: Library of Congress, 2011. (www.loc.gov/standards/premis/v2/premis21.pdf)

PURL. (code.google.com/p/persistenturls/w/list)

Reilly, Sean, and Robert Tupelo-Schneck. "Digital Object Repository Server: A Component of the Digital Object Architecture." *D-Lib Magazine* 16: 1 / 2 (January/February 2010). (www.dlib.org/dlib/january10/reilly/01reilly.html)

Tonkin, Emma. "Persistent Identifiers: Considering the Options." *Ariadne* 56 (July 2008).

3.7. File Naming

A file naming convention is a set of agreed-upon rules used to assign identifiers to digital objects in a collection. A naming convention ensures that files can be *consistently* and

uniquely identified within the repository system and is thus essential to data integrity and internal workflows. The focus of the following recommendations is digitized analog material. In the case of born-digital objects, an institutional file naming convention may also be applied, but in this case the original file name must be preserved as part of the provenance metadata.

HOPE will not take it upon itself to recommend a single naming convention or workflow for all CPs but will instead set out general guidelines to help CPs set their file naming practice and procedures.

3.7.1. File Naming: Characteristics

A good naming convention should:

- Be standardized, stable, and applicable to all collections and projects in the institution;
- Avoid identical file names to prevent accidental overwriting and loss of files;
- Enforce unambiguous distinction between files to allowing files to be easily identified (directly, through the name itself or indirectly, through a metadata record);
- Provide the means to easily distinguish among the different instances of a file (format, quality, etc.);
- Support complex digital objects-objects comprised of two or more content files of the same format and having a physical and/or logical relationship to one another;
- Facilitate the retrieval and processing of materials from creation onwards.

All file names should comply with the following minimum requirements:

Character set: Characters should be in lower case, and only alphanumeric characters should be used, with the exception of hyphen "-", underscore "_", and period "." (for the file extension). Spaces should not be used.

Length: Under normal conditions, all operating systems support file names consisting of 255 characters. It is, however, advised to restrict file names to about 30 characters, including the period "." and extension, as some operating systems are unable to handle very long paths, which can lead to copying errors.

HOPE recommends that CPs develop and use a naming scheme that is logical, consistent, and stable (i.e. not based on values which are subject to modification); does not duplicate names or values; supports complex objects and multiple derivative formats; and complies with the minimal character and length guidelines.

3.7.2. File Naming: Elements

Every file name is comprised of a few basic elements. Some are mandatory, while others are optional. These include:

1. Institutional prefix
2. Root file name

3. Sequence designator
4. Quality suffix
5. Processing suffix
6. File extension

Only the root file name and file extension elements are mandatory for every instance of a file name. The composition of the file name may vary, even when using the same naming convention, depending on the material being named. An underscore "_" should be used to separate any of the first five elements. A period should be used to separate the file extension from the other elements.
Institutional Prefix: The prefix should be a unique identifier designating the institution that created or has custody of the digital object. If possible the identifier should include a formal country code specified according to ISO 3166 and a national repository code or other unique institutional identifier. The institutional prefix is particularly helpful if material will be exchanged or aggregated with the material from other institutions.

```
Example: hu-osa
```

Root File Name: The root file name is a name given to the file to distinguish it from other files created or stored in the same institution. The name may be "descriptive", incorporating some characteristic of the content, such as its predominant content or its call number, or the name may be "non-descriptive", completely arbitrary and devoid of any reference to characteristics of the file's content.

Descriptive root file names contain words, numbers, or abbreviations that describe in some way the file they pertain to. They may be composed of a title, the name of the creator, the accession number of the physical item, subcollection or media designation or some other descriptive identifier. Meaningful root names make it easier to identify and manage the digital files and require less dependence on catalogue software, reducing the impact when something goes awry with this software. Descriptive names may also facilitate end user access to and use of material. On the downside, meaningful file names are often specific to particular collections and should be conceived for each project, so they are only feasible for medium to small collections. Furthermore, there is the added possibility that the name's meaning will be lost or change connotation over time or that the convention will not scale well as collections grow and change.

```
Example: hu-osa_mss64
```

Non-descriptive root file names express no relationship to the item and are usually sequential numbers. Non-descriptive root names work well for medium to large collections, are easy to assign and add automatically. Non-descriptive root file names provide no identifying information; thus the files are harder to manage and workflows center on the database that contains the associated metadata. The decision to use meaningful or non-descriptive file names should be based on the collection's characteristics, repository current and future requirements, and internal resources.

```
Example: hu-osa_12345678
```


HOPE recommends the use of non-descriptive numbers or codes as root file names only for medium or large collections or institutions with robust repository infrastructures. For smaller-scale collections supported by less developed infrastructure, it is advisable to use descriptive root file name, based on call numbers, local identifiers, or archival reference numbers or some combination of elements representing the intellectual structure of institutional holdings.

Sequence Designator: Files belonging to the same compound digital object (e.g. the digitized pages of a diary) should have the same root file name. In such cases, to distinguish one file from another and to indicate the relative position of one file in the sequence of files, a sequence designator should be used. The sequence designator aids in expressing the structural relationship of the files so that the digital object can be displayed in the proper sequence to an end user. The value of the sequence indicator should be a number between 1 and n, with 1 designating the first file in the sequence of files, and n designating the following files.

```
Example: hu-osa_mss64_001
```

It is important to remember to add 0s in front of the numbers to facilitate automatic sorting.

Quality Suffix: A quality suffix should be used only to distinguish different levels of quality for files of the same file format to prevent reuse of an identical file name. In this context, quality is used to indicate the richness of a file or the use to which the file will be put.

In the HOPE data model, there are five defined quality levels for any given digital object: the master, high-resolution derivative, low-resolution derivative, preview, and thumbnail. If the same file format is used, for multiple quality levels can be distinguished by adding the following quality suffixes:

- *m for the master*
- *h for a high-resolution derivative*
- *l for a low-resolution derivative*
- *p for a preview*
- *t for a thumbnail*

```
Example: hu-osa_mss64_001_h
```

Processing Suffix: If a file has been edited, and needs to be distinguished from an unedited version of the same file, this should be indicated in the filename by a lowercase "e". For example, an original file may be edited to modify the content of the file in some way, such as to delete unwanted artifacts or confidential text or to insert content. In this case, the following name might apply:

```
Example: hu-osa_mss64_001_h_e
```

File Extension: The file extension is a three-or four-letter string designating the file format. For example: *.html,*.sgml, *.tiff, *.jpg, *.gif, *.mpeg, etc. File extensions are usually generated by the software application used to create the content file.

```
Example: hu-osa_mss64_001_h_e.tiff
```

3.7.3. File Naming: Directory Conventions

Many of the rules for file names also apply for directory names. Often, the file naming is integrated with the directory structure rules, the file name replicating to some degree the structure. In this case, it is important that the file name does not depend on its location in the structure for its uniqueness but that it can function independently as a file identifier. Other than this, the directory structure should comply with the following minimum requirements:

- Restrict folder names to 30 characters
- Restrict the amount of subfolders to five (not counting the root folder).

3.7.4. File Naming: Workflow

When digitizing materials, the three possible file naming procedures are:

- automatically producing file names with scanning software;
- manually editing after scanning;
- running a script that batch renames files according to custom rules.

The choice largely rests on the broader digitization and digital curation workflows, e.g. when and how files are created; when and how quality control is undertaken; when and how files are packaged into objects; when and how file names are stored in the local system; when and how derivatives are created; when and how objects are stored in the file server or on storage media; etc. As a rule, manual editing is discouraged as it is labor intensive and prone to human error. In any case, naming conventions should be agreed upon and documented in advance of digitization-and not applied retrospectively. Policies should be set indicating whether naming conventions are project or collection based or institution wide. The latter is preferable, if only because it is more scalable, reducing the risk of confusion in the long term.

HOPE recommends that all CPs create a clear file and directory naming convention, if possible one which applies across all collections and projects. General workflows should be developed which integrate and support this naming convention.

3.7.5. File Naming in HOPE: Recommendations

For the **SOR**, it is recommended that unique file names be generated upon upload of master objects. A standard SOR file naming convention will facilitate the creation of derivatives as well as later migration. We recommend SOR master file names be composed of the following elements:

- *Institutional Prefix*, based on an ISO country code and unique repository code for each CP;
- *Root File Name*, based on the file's PID;
- *Quality Suffix*, if applicable, based on the HOPE-defined quality levels;
- *File Extension*.

In this case, originally submitted master file names should be preserved as part of the provenance metadata.

For **LORs**, it is recommended to set up local file and directory naming conventions which can serve to uniquely identify the file and to preserve aspects of its provenance. The use of non-descriptive file root names eases the generation of file names but can also cause problems if the link between the objects and their metadata ever breaks. They are therefore recommended for medium or large institutional collections or those with a robust repository infrastructure. Otherwise, it is advisable to store as much information as possible in the file name without complicating workflows or relying too heavily on manual entry. (Note that even information rich file names can be produced (semi)automatically.)

In general, it is good practice to document the institutional file naming conventions to ensure the same rules are used with every digitization project. Importantly, rules should not be applied retroactively to existing content but should rather serve and the basis for future digitization projects. If whether by fault or by design, file names are changed, it is recommended to store old file names as provenance information.

Finally, CPs should avoid using file names and directory structures as their sole structural metadata but should instead attempt to store structural metadata in a more robust manner. HOPE recommends the use of METS to capture the structural metadata on digital objects. The HOPE System imposes no additional requirements regarding local file names. Beyond the minimal technical requirements, file naming conventions should be developed to suit local needs, and file naming procedures should be integrated into digitization, transformation, processing, and storage workflows.

3.7.6. File Naming: References

State Library of Queensland. *Directory & File Naming Conventions for Digital Objects, v1.06.* 2012.
(linked from Queensland Government site: www.slq.qld.gov.au/about/pol)

UCSD Libraries Digital Library Program. *A Naming Protocol for Digital Content Files*. 2003.
(libraries.ucsd.edu/artsnet/fvlnet/filename_conventions.pdf)

University Library, University of Illinois at UrbanaChampagne. *Library Digital Content Creation: Best Practices for File Naming*. 2010.
(www.library.illinois.edu/dcc/bestpractices/chapter_02_filenaming.html)

3.8. Technical Metadata

Although technical metadata is only a subset of the complete suite of administrative metadata necessary to manage, secure, and provide access to digital objects, it has often been called the first line of defense. Technical metadata assures that the information content of a digital file can be resurrected even if traditional viewing applications associated with the file have vanished. Furthermore, it provides metrics that allow machines, as well as humans, to evaluate the accuracy of output from a digital file. In its entirety, technical metadata supports the management and preservation of digital content through the different stages of its life cycle.

3.8.1. Technical Metadata: Selecting Standards

Currently, there exists no "out of the box" technical metadata standard suitable for all kinds of digital materials. Most available technical metadata standards were created and finalized years ago or have remained in an early "beta" stage-perhaps indefinitely. PREMIS, last updated in 2011, is a continuously evolving model which defines itself as a common subset of all the metadata needed by an organization running a preservation repository. PREMIS covers a broad range of metadata on rights, events, and agents as well as digital objects. Only a subset of the PREMIS semantic units describing the Digital Object entity can be considered technical. That being said, PREMIS can provide a core set of technical metadata to be extended by more particular media-specific standards.

Media-specific technical standards tend to be exhaustive-attempting to identify all possible elements that might characterize the digital object-as they are created for a range of environments and purposes. For this reason, the technical standards should not be used as strict guidelines but should be regarded as a set of options from which to choose. When selecting media-specific elements, it is important to consider:

- the nature of the digital collections;
- the needs and requirements of the repository's target users;
- the functions that the repository will be asked to fulfill;
- the feasibility and method of collecting and storing the metadata.

As the HOPE System does not yet provide full preservation services, only a minimum set of elements are recommended. HOPE recommends that CPs define a core set of PREMIS technical elements with media-specific extensions that can be more or less exhaustive depending on local needs and resources and the nature of the digital content.

3.8.2. Technical Metadata in PREMIS

The following is a list of technical elements defined by PREMIS to describe the Digital Object entity. PREMIS limits the scope of its work to elements that would apply across all formats. These can be seen as essential elements, the collection of which should be prioritized in local workflows.

- *objectCharacteristics*: Technical properties of a file or bitstream that are applicable to all or most formats.

  - *compositionLevel*: An indication of whether the object is subject to one or more processes of decoding or unbundling.
  - *fixity*: Information used to verify whether an object has been altered in an undocumented or unauthorized way. (See: Section 8 Fixity).
  - *size*: The size in bytes of the file or bitstream stored in the repository.
  - *format*: Identification of the format of a file or bitstream where format is the organization of digital information according to preset specifications.
  - *creatingApplication*: Information about the application that created the object.
  - *inhibitors*: Features of the object intended to inhibit access, use, or migration.
  - *objectCharacteristicsExtension*: A container to include semantic units defined outside of PREMIS.

- *environment*: Hardware/software combinations supporting use of the object.
  - environmentCharacteristic: An assessment of the extent to which the described environment supports its purpose.
  - environmentPurpose: The use(s) supported by the specified environment.
  - environmentNote: Additional information about the environment.
  - dependency: Information about a non-software component or associated file needed in order to use or render the representation or file, for example, a schema, a DTD, or an entity file declaration.
  - software: Software required to render or use the object.
  - hardware: Hardware required to render or use the object.
  - *environmentExtension*: A container to include semantic units defined outside of PREMIS.

- *signatureInformation*: A container for PREMIS defined and externally defined digital signature information, used to authenticate the signer of an object and/or the information contained in the object.
  - signature: Information needed to use a digital signature to authenticate the signer of an object and/or the information contained in the object.
  - signatureInformationExtension: Digital signature information using semantic units defined outside of PREMIS.

PREMIS can be extended using "Extension" containers: objectCharacteristicsExtention, environmentExtension, signatureInformationExtension. The PREMIS working group suggests that when you extend PREMIS, you observe the following principles:

- An extension container may be used to either supplement or replace PREMIS semantic units within the parent container. The one exception is objectCharacteristicsExtension, which may only supplement objectCharacteristics.
- An extension container may be used with existing PREMIS semantic units, supplementing the PREMIS semantic units with additional metadata.
- An extension container may be used without existing PREMIS semantic units, effectively replacing the PREMIS semantic units with other applicable metadata.
- Where there is a onetoone mapping between the contents of an extension container and an existing PREMIS semantic unit, recommended best practice would be to use the PREMIS semantic unit rather than its equivalent in the

extension; however, implementers may choose to use the extension alone, if circumstances warrant.
- If any semantic unit is not used it should be omitted, rather than an empty schema element included.
- If the information in an extension container needs to be associated explicitly with a PREMIS unit the parent container is repeated with appropriate subunit. If extensions from different external schemas are needed, the parent container should also be repeated. In this case the repeated parent container may include the extension container with or without any other existing PREMIS semantic units for that parent container.
- When an extension container is used, the external schema being used within that extension container must be declared.

(See Appendix B for Media Specific Standards)

3.8.3. Technical Metadata: Collection and Storage Workflows

Unlike descriptive metadata, technical metadata must be collected from different sources over the entire course of an object's life cycle. Thus, robust workflows for the collection and short- and long-term storage of technical metadata are essential. When setting up technical metadata workflows, it is important to consider the following:

- *Source*: how is the metadata created and at what point can the metadata be captured?

Examples:
It is intellectual information that can only be gathered manually at the point of creation, e.g. hardware or software information.
The file itself carries the information, but it is not possible to extract it, e.g. low-level codec information.
The file itself carries this information and could be extracted with a file validation tool, or could be generated automatically, e.g. mime-type, dimension, color-depth.

- *Storage*: how should the data be stored over the short and long term?

Examples:
The data is stored in a database with other object metadata.
The data is recorded by digitization vendor in an Excel sheet or machine readable form and can be imported into a management system when needed.
The data is embedded in the file and can be extracted when needed.
The data is known by staff and can be elicited when needed.

- *Basis*: the technical metadata is applicable to the object in which version(s) or form(s)?

Examples: master, derivative, born-digital master, raster image formats, etc.

- Granularity: at which level should the metadata be captured?

Examples: bitstream, file, object, object group, collection.

As the HOPE System does not yet provide full preservation services, a relatively light workflow is recommended to support the collection and storage of technical metadata. CPs should focus on gathering the metadata which cannot be gathered at a later point and storing it in a machine readable form. Only those elements which are needed for the daily functioning of local and HOPE systems are necessary to be stored and exportable in XML form. HOPE itself currently requires little technical metadata. Those CPs using the SOR are currently required to submit master File Format information along with the digital object.

3.8.4. Technical Metadata in HOPE: Recommendations

As mentioned in Administrative Metadata, the **Aggregator** has as its primary function to store and disseminate descriptive packages and their related Dissemination Information Packages - in other words to deliver objects in a form that can be rendered in the online environment and understood and used by our target users. For this, it is recommended that the Aggregator maintain sufficient representation information to allow a digital derivative object to be accessed, rendered, and used by our target user group. The metadata recommended below should be submitted to the Aggregator.

Highly recommended, metadata on:

- File format and size of access derivative, file level.

Recommended, metadata on:

- Physical characteristics which inhibit access on access derivatives or masters, *description or object level*;
- Access facilitators (e.g. time coding) on access derivatives or masters, *file level*.

As mentioned, the **SOR** has as its primary function to ingest, store, and make available over the medium term digital master objects as well as to support object transformation (i.e. derivative creation). For this, it is recommended that the SOR store information necessary to create the derivatives in a format required by portals along with other representation information needed to make the content usable by our target users. The fixity information to support routine quality control is also important. The metadata recommended below may be submitted to or generated by the SOR, or some combination.

Highly recommended, metadata on:

- Fixity of masters, *file level*;
- Viewers and players that are not readily available to the average user, specifically AV players (links to external viewers or embed links would also suffice) for masters and derivatives, *file or object level*;
- File format of masters and derivatives, *file level*;
- File size for masters and derivatives, *file level*.

Recommended, metadata on:

- Format version and registry information for masters, *file level*;
- Fixity of derivatives, *file level*.

As mentioned, **LORs** generally serve a range of functions. In all cases, LORs need to be able to produce sufficient representation information to represent objects online for target users; metadata supporting this function should be prioritized. In the case of non-SOR users, they may also support ingest and storage of master files and derivative creation as well as routine quality control; metadata supporting this should also be collected. In general, we advise that LORs support recommended elements (See Appendix C for media type elements) in a robust data management system. However, Like PREMIS, our policy is that LORs should be able to produce recommended metadata when needed. Currently, LORs are only required to produce file format information on submitted files. Beyond this, we make no further presumptions.

3.8.5. Technical Metadata: References

Caplan, Priscilla. *Understanding PREMIS*. Washington D.C.: Library of Congress, 2009. (www.loc.gov/standards/premis/understandingpremis.pdf)

Library of Congress. *AudioMD Data Dictionary*. (www.loc.gov/rr/mopic/avprot/DD_AMD.html)

Library of Congress. *VideoMD Data Dictionary*. (www.loc.gov/rr/mopic/avprot/DD_VMD.html)

National Library of New Zealand. *Metadata Standards Framework-Preservation Metadata (Revised)*. Wellington, N.Z.: NLNZ, 2003. (www.natlib.govt.nz/downloads/metaschemarevised.pdf)

Network Development and MARC Standards Office, Library of Congress. *NISO Metadata for Images in XML Schema*. (www.loc.gov/standards/mix/)

Network Development and MARC Standards Office, Library of Congress. *TextMD: Technical Metadata for Text*. (www.loc.gov/standards/textMD/)

NISO. *NISO Standard Z39.87, Technical Metadata for Digital Still Images*. (www.niso.org/apps/group_public/project/details.php?project_id=69)

OCLC/RLG Working Group on Preservation Metadata. *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*. Dublin, OH: OCLC, 2002. (www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf)

PREMIS. *PREMIS Data Dictionary for Preservation Metadata, v. 2.1*. Washington D.C.: Library of Congress, 2011. (www.loc.gov/standards/premis/v2/premis21.pdf)

3.9. Fixity

Fixity, in preservation terms, means that the digital object has not been changed between two points in time or two events. Technologies such as checksums, message digests, and digital signatures are used to verify a digital object's fixity. Fixity information, the information created by these fixity checks, provides evidence on the bit integrity of the digital objects and is thus an essential element of a trusted repository.
A fixity check may be used to verify that any action taken upon the digital resource does not alter the resource. Fixity checks all work in the same basic way: a value is initially generated and saved; it is then recomputed and compared to the original to ensure that the object (file or bitstream) has not changed.

PREMIS distinguishes fixity information from digital signatures, which are used to guarantee the authenticity of the object and are created by the document producer, submitter, or even the archive itself to connect the agent with the object. Digital signatures are unique to the signature producer, but they also relate to the content of the document -the process of creating and verifying digital signatures relies on the generation and checking of fixity values generated using a Secure Hash Algorithm (SHA). Both the creator of the signature and the fixity of the document from the point that the signature was created are needed to confirm a document's authenticity.

3.9.1. Fixity: Checksums

A checksum is the simplest yet least secure method of verifying fixity. Checksums are typically used in error-detection to find accidental problems in transmission and storage. The least complicated checksum algorithms do not account for such changes as the reordering of bytes or changes that cancel one another out. The more secure checksums, such as cyclic redundancy check (CRC) are hash functions that control for such changes. Because of the comparative simplicity of their mathematical algorithms, however, checksums are vulnerable to deliberate and malicious tampering.

CRC Cyclic redundancy check: the cyclic redundancy check, or CRC, is a technique for detecting errors in digital data, but not for making corrections when errors are detected. It is used primarily in data transmission. In the CRC method, a certain number of check bits, often called a checksum, are appended to the message being transmitted. The receiver can determine whether or not the check bits agree with the data, to ascertain with a certain degree of probability whether or not an error occurred in transmission. If an error occurred, the receiver sends a "negative acknowledgement" (NAK) back to the sender, requesting that the message be retransmitted.

3.9.2. Fixity: Message Digest Algorithms

Unlike checksums, cryptographic hash functions such as message digests are not prone to attack. A message digest is computed by applying an algorithm to the file of any length to produce a unique, short, uniform length character string. What makes message digests more secure than checksums is the complexity of the algorithm. A message digest is like the fingerprint of a digital object. Hashes are oneway operations; a hash can be created from a digital object, but the digital object cannot be recreated from the hash. MD5 and Secure Hash Algorithm, SHA1, are commonly used cryptographic hash algorithms.

MD5 Message Digest Algorithm 5: The MD5 algorithm takes as input a message of arbitrary length and produces as output a 128bit "fingerprint" or "message digest" of the input. It is conjectured that it is computationally infeasible to produce two messages having the same message digest, or to produce any message having a given pre-specified target message digest. The MD5 algorithm is intended for digital signature applications, where a large file must be "compressed" in a secure manner before being encrypted with a private (secret) key under a public-key cryptosystem such as RSA. In essence, MD5 was a way to verify data integrity, and is much more reliable than checksum and many other commonly used methods. MD5 is a widely used algorithm and is supported by many programming APIs currently in use.

SHA1 Secure Hash Standard: SHA is a cryptographic message digest algorithm similar to the MD4 family of hash functions developed by Rivest. It differs in that it adds an additional expansion operation, an extra round and the whole transformation was designed to accommodate the DSS block size for efficiency. The Secure Hash Algorithm takes a message of less than 264 bits in length and produces a 160bit message digest which is designed so that it should be computationally expensive to find a text which matches a given hash.

The HOPE SOR currently uses the MD5 algorithm for ingest and routine fixity checks.

3.9.3. Fixity: Digital Signatures

Digital signatures combine a hash message digest with encryption. A digital signature starts with the creation of a message digest from the digital object. The message digest is then encrypted using asymmetric cryptography. Asymmetric cryptography uses a pair of keys: a private key to encrypt and a public key to decrypt. The private key must be held secretly and securely by the signer. The signature can be verified by decrypting the signature with the signer's public key and comparing the now-decrypted digest with a new digest produced by the same algorithm from the same content.

A reliable digital signature requires that:

- The process of producing a signature is considered to be unique to the producer.
- The signature is related to the content of the document that was signed.
- The signature can be recognized by others to be the signature of the person or entity that produced it.

As the PREMIS report outlines, digital signatures are used in preservation repositories in three ways:

- For submission to the repository, an agent (author or submitter) might sign an object to assert that it truly is the author or submitter.
- For dissemination from the repository, the repository may sign an object to assert that it truly is the source of the dissemination.
- For archival storage, a repository may sign an object so that it will be possible to confirm the origin and integrity of the data. In this case, the signature itself and the information needed to validate the signature must be preserved.

HOPE does not yet require or support the use of digital signatures.

3.9.4. Fixity: Workflows

General workflows for generating, storing, checking fixity information include:

- At least one but ideally more fixity values should be generated at the point of digitization or ingest. The algorithm used should always be preserved along with the fixity value.
- Each fixity check should be treated as a digital lifecycle event. Fixity checks should be tracked, for instance in audit trail or log file; information recorded should include a date/time stamp, staff member performing the check, and result of the fixity check.
- The generation of new or additional fixity values should be regarded as a digital life cycle event. Fixity value generation should be tracked, for instance as part of an audit trail or log file; information recorded should include the date/time stamp and staff member generating the value.
- Best practices within the digital preservation community regarding fixity checks should be continually reviewed. Particular attention should be paid to the use of digital signatures.

3.9.5. Fixity in HOPE: Recommendations

It is *not recommended* that the **Aggregator** collect, generate, or store fixity information.

It is *highly recommended* that the **SOR**:

- Support at least one common fixity algorithm, either a message digest or digital signature;
- Collect and store locally generated fixity values (along with the fixity algorithm used) for each submitted master file;
- Generate an independent fixity value for each submitted digital master file during ingest and run fixity check by comparing with locally generated value;
- Run routine and event-driven fixity checks to ensure the continuing integrity of the master files;
- Generate new fixity values when transformation is performed;
- Enable export of fixity values and algorithms along with master files;
- Store event information on fixity generation, checks, and export of values;

- Support the update of fixity check method and algorithm.

It is *highly recommended* that **LORs**:

- Support more than one common fixity algorithm/method; if CP is using the SOR, then this would include the MD5 algorithm currently supported by the SOR;
- Generate at least one fixity value for each master and store along with algorithm used;
- Enable the export or delivery of fixity values for each master file to SOR or other systems, as needed;
- Manage and update fixity values during life-cycle events, i.e. migration, transfer, export of master files;
- Store basic event information on the generation of fixity values, fixity checks, fixity value exports, and fixity value updates;
- Support the update of fixity check method and algorithm.

3.9.6. Fixity: References

National Institute of Standards and Technology (NIST). *Federal Information Processing Standards Publication: Secure Hash Standard (SHS)*. Gaithersburg, MD: NIST, 2008.
(csrc.nist.gov/publications/fips/fips1803/fips1803_final.pdf)

Network Working Group. *RFC 1321: The MD5 MessageDigest Algorithm*. 1992.
(www.ietf.org/rfc/rfc1321.txt)

Novak, Audrey (ILTS). *Fixity Checks: Checksums, Message Digests and Digital Signature*. 2006.
(www.library.yale.edu/iac/DPC/AN_DPC_FixityChecksFinal11.pdf)

PREMIS. *PREMIS Data Dictionary for Preservation Metadata, v. 2.1*. Washington D.C.: Library of Congress, 2011.
(www.loc.gov/standards/premis/v2/premis21.pdf)

Case Study: Open Society Archives (OSA) Manages Administrative Metadata

Since it opened its doors in 1995, the Open Society Archives (OSA) at the Central European University in Budapest (Hungary) has collected and made openly accessible material on communism, the Cold War, and their aftermath and on human rights worldwide. OSA holds approximately 7,000 linear meters of archival records (including audio visual content) and a library collection which comprises more than 6,500 dailies, journals, and informal press titles. OSA has approximately 30 staff members with a small dedicated IT unit, which works in coordination with the university's main IT unit. Prior to the project, OSA depended on several small in-house developed solutions (separate for archives, library, and film library) to catalogue their collections and make them available on their website. Separate databases were likewise created for each digitized collection. For OSA, the HOPE project was primarily a means to improve the level of their internal systems and practices-in particular they were keen to introduce a robust digital object repository. OSA opted to join the HOPE PID Service and the SOR, hoping to work closely to integrate these systems into their envisioned repository.

It was clear from the outset of the HOPE project that not only was a digital object repository a desirable outcome, but it would also be necessary in order to meet HOPE requirements. The existing system was fragmented with a separate data structure for each digital project and little or no administrative metadata kept on the digitized objects. The storage of digital objects was also idiosyncratic,with masters generally stored on tapes and derivatives distributed over several servers or stored directly in the website. OSA's first step was to develop a common metadata schema, which included both descriptive metadata (for archival and library items and collections) and technical metadata on the related digital files. The schema was based on known standards, primarily EAD, MARC, and PREMIS. (For archival description, OSA opted to define their own elements based on EAD rather than to directly use the schema; MARC and PREMIS elements were directly incorporated.) A new architecture was introduced based on:

- Fedora Commons: low-level metadata storage;
- Apache Solr: search engine;
- Drupal: website CMS and content display UI.

Fedora is designed around "compound digital objects", whereby one or more "content items" are aggregated into the same digital object. Content items can be of any format and can either be stored locally in the repository or stored externally and referenced by the digital object . Each content item in Fedora is represented by a datastream. OSA found Fedora to have "excellent flexibility", allowing them to design objects according to institutional needs. In the end, OSA decided on an atomistic "everything is an object" approach and defined objects in the following way:

Collection objects include a descriptive metadata datastream for a single collection;

Item objects include:
- Descriptive metadata datastream for a single item (multipage document, single-page document, film, etc.)
- Relationship metadata on the related collection
- METS metadata datastream to relate file objects to items

File objects include:
- File (master, derivative, or other), as externally managed content on OSA's file server or in the SOR
- Technical metadata, as an externally managed XML file on OSA's file server

Each collection and item object will have a unique ID in the form of "osa:" followed by a 32bit hex GUID hash. Each file object will have an ID in the same form as the ID of the item to which it belongs followed by a quality suffix and sequential page information. These IDs will form the root of the OSA Handles stored by the HOPE PID Service.

For example:

Fedora item ID: osa:3b34347820ef45f0bfd87239c096c5a1
Item PID: hdl:10891/osa:3b34347820ef45f0bfd87239c096c5a1
http://hdl.handle.net/10891/osa:3b34347820ef45f0bfd87239c096c5a1
Fedora file ID: osa:3b34347820ef45f0bfd87239c096c5a1_m_0001
(from above item, first page of master file)

Master file PID: (location attribute assigned by SOR)	hdl:10891/osa:3b34347820ef45f0bfd87239c096c5a1_m_0001 http://hdl.handle.net/10891/ osa:3b34347820ef45f0bfd87239c096c5a1_m_0001? locatt=view:master
Thumbnail file PID: (thumbnail generated and location attribute assigned by SOR)	hdl:10891/osa:3b34347820ef45f0bfd87239c096c5a1_m_0001 http://hdl.handle.net/10891/ osa:3b34347820ef45f0bfd87239c096c5a1_m_0001? locatt=view:thumbnail

(File names follow the same convention as Fedora file IDs.) The ID system went through several iterations as OSA adapted and defined its workflow and data structure. The final convention was determined in a meeting involving professional colleagues, IT, and management-a meeting which also treated archival reference codes and library special collection call numbers. The need to create a permanent and lasting PID convention drove this process. The IDs and names thus arrived at are surely robust and also reflect the relationship of entities in OSA's system. On the other hand, they exceed recommended lengths, which may hinder internal administration and possibly external use. The inclusion of an institutional acronym could also prove a problem over the long term.

In addition to local IDs and PIDs, OSA plan to capture and store a range of technical metadata, including provenance information. As mentioned, OSA use PREMIS as a base schema but include extensions to other format-specific schema, e.g. NISO MIX, videoMD, and audioMD; PREMIS nicely accommodates extensions. To automatically capture as much of this metadata as possible, OSA tested several available solutions across various content formats comparing their results to their technical metadata requirements. In the end, OSA have opted for two solutions: Jhove (for documents and images); MediaInfo (for audio and video files).
These were chosen based on "the completeness of results and their active development status". Both programs generate XML files as an output. These technical metadata files will be named in accordance with master copies and stored in the same directory structure on a local file server. To facilitate automatic metadata capture, OSA plan to develop several small applications. The first will check the directory structure to locate any master files without an accompanying technical metadata XML file; if located the application will check the mime type and trigger the generation of metadata using either Jhove or MediaInfo. The second will create a datastream pointing to the technical metadata XML file, as part of the Fedora file object creation process. Fedora will be able to pull in the content of the technical metadata file upon request.

Several values cannot be generated automatically during this process, and OSA are still looking for possible solutions to this problem. These include metadata on the hardware and software used to create the files as well as the creation date, format registry information, and original file names. OSA plan to look into possibilities for embedding more metadata into files at the point of digitization, using metadata schemas such as exif or xmp. They may also consider creating a collection level technical metadata datastream to store global values such as the names of service providers or format registry information as well as external links to scanning logs and file naming tables on the whole collection. They point out that since this information is primarily for long-term preservation, it is currently unnecessary to store it at a high level of granularity.

Beyond this OSA still plan to develop modules for Rights and Events management, also based on the PREMIS model, though extended to meet local needs. Events will allow OSA to track new file creation and deletion, fixity checks, and other information related to the objects history after submission to the repository. Rights will be more problematic. OSA are currently looking into methods which will allow them to manage the rights for all their holdings-analog and digital-across the entire archival workflow.

OSA were deeply involved in repository best practice work and decided to use their own in-house development as a test-bed for practices they were advocating. Thus far, OSA have implemented many HOPE best practice recommendations into their data structure and system and have developed several solutions to generate, store, and manage administrative metadata. Nevertheless, OSA are still at the beginning of a long process, and the real test will come as they attempt to manage administrative metadata through the entire process of digitization, ingest, storage, harvesting, and migration.

CONCLUSION

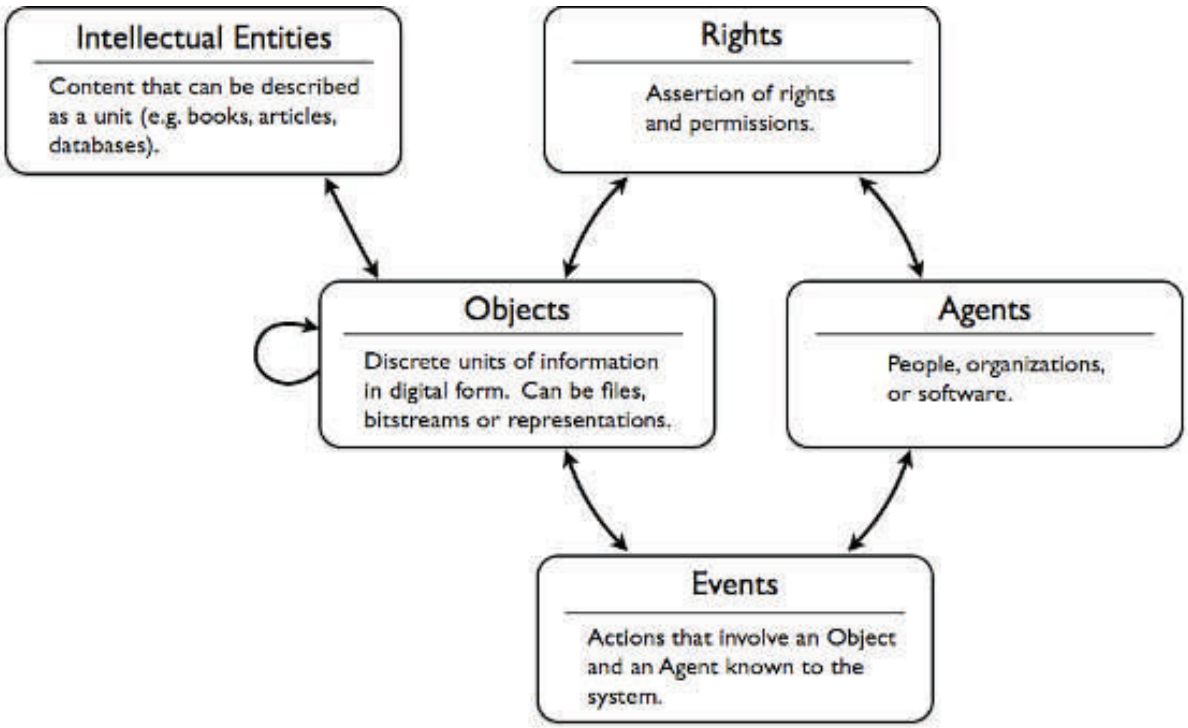
Today HOPE partners feel a pressing need to populate their websites with vast amounts of content and to actively push their content through innovative channels to targeted users and the broader public. Data aggregation has been a strong trend for decades now, satisfying the thirst for content and fitting well into policy agendas and technical trends. Many social history institutions see long-term preservation as a conflicting priority. But perhaps without realizing it, HOPE CPs already support some level of preservation: accepting donations (even digital donations), gathering metadata, applying professional standards, maintaining a policy framework to ensure a certain level of service. By setting out best practices informed by the OAIS model and Trusted Digital Repository guidelines we have attempted to demonstrate that, even in the short term, there is a lot to gain from digital curation: to identify threats and losses; to demonstrate sustainability and viability; to enhance trust in archival institutions; to promote transparency, open access, and open standards.

This task has intended to draw attention to the importance of reliability and trustworthiness in our sector, begging that these be considered alongside other priorities. As HOPE was primarily conceived to enhance the visibility of social history sources through different services, we have at all times refrained from prescribing strict measures. Instead, we have attempted to offer nuanced risk evaluation and best practice recommendations with the goal of fostering discussion about the meaning and role of "the trusted digital repository" in the HOPE context.

A. APPENDIX - PREMIS and NZLZ

A.1 PREMIS: Data Model

The PREMIS Data Model comprises five entities: *Objects*, *Events*, *Agents*, *Rights*, and *Intellectual Entities*.



A-1. Diagram – PREMIS Data Model

Objects and *Agents* exist as the two primary nodes, connected to each other through *Events* and *Rights*.

The **Objects** are what are actually stored and managed in the preservation repository. Most of PREMIS is devoted to describing digital objects, concentrating primarily on its technical characteristics. The information that can be recorded includes:

- a unique identifier for the object (type and value),
- fixity information such as a checksum (message digest) and the algorithm used to derive it,
- the size of the object,
- the format of the object, which can be specified directly or by linking to a format registry,
- the original name of the object,
- information about its creation,

- information about inhibitors,
- information about its significant properties,
- information about its environment,
- where and on what medium it is stored,
- digital signature information,
- relationships with other objects and other types of entities.

PREMIS defines three different kinds of objects and requires implementers to make a distinction between them. These are file objects (the primary unit), representation objects (which are made up of file objects), and bitstream objects (which make up file objects). Some semantic units defined in the PREMIS Data Dictionary are applicable to all three types of object, while others are applicable to only one or two types of object.

The **Event** entity aggregates information about actions that affect objects in the repository. An accurate and trustworthy record of events is critical for maintaining the digital provenance of an object, which in turn is important in demonstrating the authenticity of the object. The information that can be recorded about events includes:

- a unique identifier for the event (type and value),
- the type of event (creation, ingestion, migration, etc.),
- the date and time the event occurred,
- a detailed description of the event,
- a coded outcome of the event,
- a more detailed description of the outcome,
- agents involved in the event and their roles, • objects involved in the event and their roles.

The Data Dictionary entry for Type provides a "starter list" of events to help guide implementation.

Agents can be people, organizations, or software applications. PREMIS defines only a minimum number of semantic units necessary to identify agents, since there are several external standards that can be used to record more detailed information. A repository could choose to use a separate standard for recording additional information about agents, or it could use the agent identifier to point to externally recorded information. The Data Dictionary includes:

- a unique identifier for the agent (type and value),
- the agent's name,
- designation of the type of agent (person, organization, software).

The **Rights** entity aggregates information about rights and permissions that are directly relevant to preserving objects in the repository. Each PREMIS rights statement asserts two things: acts that the repository has a right to perform, and the basis for claiming that right. The information that can be recorded in a rights statement includes:

- a unique identifier for the rights statement (type and value),
- whether the basis for claiming the right is copyright, license or statute,

- more detailed information about the copyright status, license terms, or statute, as applicable,
- the action(s) that the rights statement allows,
- any restrictions on the action(s),
- the term of grant, or time period in which the statement applies,
- the object(s) to which the statement applies,
- agents involved in the rights statement and their roles.

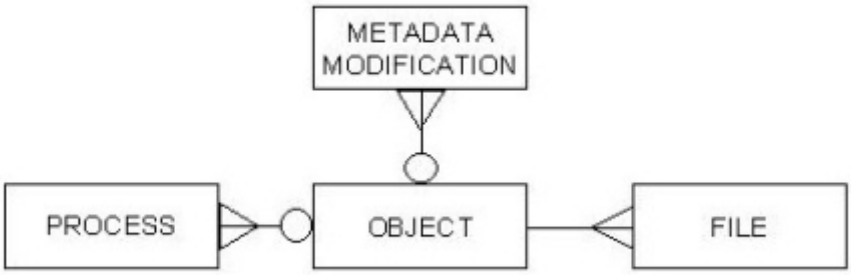
Most of the information is designed to be actionable (that is, recorded in a controlled form that can be acted upon by computer program).

The Intellectual Entities connect directly to the object. They are conceptual, and might be called "bibliographic entities." PREMIS defines an Intellectual Entity as "a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database." PREMIS does not actually define any metadata pertaining to Intellectual Entities because there are plenty of descriptive metadata standards to choose from.

A.2 NLNZ

In 2002-2003, the National Library of New Zealand developed their own preservation metadata schema as a working tool for the collection of preservation metadata applicable to its digital collections. Created in the wake of the release of OAIS reference model and during the flurry of work which would eventually produce PREMIS, the schema was designed to parallel the NISO.Z39.87 technical metadata standard for raster images. New Zealand soon after released the National Library of New Zealand Preservation Metadata Extract Tool, which complements this framework. Though the NLNZ standard has been all but supplanted by PREMIS, the extraction tool is still widely used. Today the NLNZ extraction tool can now generate technical metadata for PREMIS objects encoded using the PREMIS XML schema.

A.2.1 NLNZ High-Level Relational Data Model



A-2. Diagram - NLNZ Relational Data Model

Object contains 18 elements describing the logical object, which may exist as a file or aggregation of associated files. These elements identify the object and describe those characteristics relevant to preservation management.

Process contains 13 elements that record the complete history of actions performed on the objects. It includes the objectives of a process, who has given permission for the process, critical equipment used, and the outcomes of the actions taken. An audit trail of date/time stamps and responsible persons and/or agencies is constructed.

File contains technical information about the characteristics of each of the files that comprise the logical object identified in Entity 1. Nine elements are common to all file types, and further elements are specified for certain categories of file (e.g., image, audio, video, text).

Metadata modification contains 5 elements and records information about the history of changes made to the preservation metadata. This acknowledges that the record is itself an important body of data that must be secure and managed over time.

The NLNZ model specifies the following relationship rules:

- An Object may have one or more Processes associated with it
- An Object may have one or more Metadata Modifications associated with it
- An Object must have one or more Files associated with it
- A Process must always be associated with a single Object
- A Metadata Modification must always be associated with a single Object
- A File must always be associated with a single Object

A.3 A Comparison of PREMIS and NLNZ

One of the primary differences between PREMIS and NLNZ is that the NLNZ schema is predicated on the idea of a Preservation Master. In practice this means that various other manifestations, e.g. dissemination formats, are not considered preservation objects and will not have preservation metadata retained about them. While in PREMIS each transformation produces a wholly new object with a wholly new set of object metadata, in NLNZ, preservation masters themselves are dynamic and will be subject to further preservation processes, e.g. migration from an obsolete to a current format. This creates a life cycle of creation, use and eventual replacement and object metadata is continuously modified to reflect this -in fact, Metadata Modifications are themselves considered an entity in the model. In NLNZ, at any given time there can be only one preservation master for an object and any object carrying the status of preservation master will be subject to the maximum preservation effort whilst it has that status.

A second difference is that the NLNZ model highlights the specific structural relationship between files and objects. Preservation metadata is considered to belong to the master object in a 1:1 relationship and only indirectly to the files. Processes and Metadata Modifications are associated with objects only. The following is a list of types of digital objects defined:

- Simple objects: One file intended to be viewed as a single object (e.g., a word-processed document comprising one essay).
- Complex objects: A group of dependent files intended to be viewed as a single object (e.g., a website or an object created as more than one file, such as a database), which may not function without all files being present in the right place.
- Object groups: A group of files not dependent on each other in the manner of a complex object (e.g., a group of 100 letters originally acquired on a floppy disk). This object may be broken up into (described as) 100 single objects or 4 discrete objects containing 25 letters each, or it may be kept together as a single logical object ("Joe Blogg's Letters").

In some sense, PREMIS is granularity agnostic. Many elements can apply equally to all objects: representation objects, file objects, or bitstream objects. PREMIS offers the Object Relationship semantic unit to structure these.

Finally, NLNZ does not emphasize agents or rights. Rights, as such, are bypassed altogether, though the Process entity does have a "permissions" element. Agents are folded into their respective entities but feature particularly prominently in the Process entity.

In HOPE, the NLNZ standard can help guide in the selection of mediaspecific metadata from technical standards. It also serves to highlight the importance of distinguishing between masters and derivatives and between files, objects, and collections when selecting, creating, and storing preservation metadata.

B. APPENDIX - Technical Metadata: Media Specific Standards

The following are the commonly accepted standards for still images, text, audio, and video file formats.

B.1 NISO Standard Z39.87: Technical Metadata for Digital Still Images

This standard defines a set of metadata elements for raster images only. It does not address other image formats (e.g. vector, animated raster, motion picture). The elements document digital images created through digital photography or scanning, as well as those that have been altered through editing or image transformation. Early versions of the document referred to images maintained in TIFF. The most recent version of the standard has been expanded to include other raster image file formats. The dictionary has been designed to facilitate interoperability between systems, services, and software as well as to support the long-term management of and continuing access to digital image collections. Use of the data dictionary is accomplished primarily through XML encoding. The metadata describes the entire file (including header and other information) rather than the bitstream level.

There are four sections of the data dictionary:

- *Basic Digital Object Information*: Contains a cluster of data elements which apply to all digital object files, not just digital image files. This kind of information may be considered more general preservation metadata.
- *Basic Image Information*: The items in this section are fundamental to the reconstruction of the digital object as a viewable image on electronic interfaced displays.
- *Image Capture Metadata*: This section can best be described as descriptive technical metadata or administrative metadata. Some of the information may be harvested from the file itself while other information will need to be provided by the institution managing the image capture process.
- *Image Assessment Metadata*: The operative principle in this section is to maintain the attributes of the image inherent to its quality. These elements serve as metrics to assess the accuracy of output (today's use) and of preservation techniques, particularly migration (future use).

Although Z39.87 itself was designed to be agnostic in terms of implementation, the NISO Metadata for Images in XML Schema (MIX), commissioned by NISO and created by the Library of Congress, has been the dominant form of use for the data dictionary. Because MIX is a METS extension schema, implementation and use of the data dictionary on a local level has been fairly easy to manage.

Care has been taken to ensure that NISO Z39.87 harmonizes with PREMIS.

B.2 TextMD: Technical Metadata for Text

TextMD is a XML Schema that details technical metadata for text-based digital objects (i.e. born-digital text objects). It most commonly serves as an extension schema used within the METS administrative metadata section. However, it could also exist as a standalone document.

The textMD schema allows for detailing properties such as:

- encoding information (quality, platform, software, agent)
- character information (character set and size, byte order and size, line terminators)
- languages
- fonts
- markup information
- processing and textual notes
- technical requirements for printing and viewing
- page ordering and sequencing

B.3 AudioMD: Audio Technical Metadata Extension Schema

AudioMD is a XML schema to describe the technical characteristics of digital audio archival objects.

AudioMD contains five top level elements:

1. *bits_per_sample:* Number of bits in a digital audio sample i.e. quantization, e.g. 16, 24;
2. *channel:* Number and information about channels/tracks, e.g., 2trk, 4trk, 8trk, etc.;
3. *data_rate:* Information about the mode and data rate of audio files in Kb/s, e.g. 16, 44.1, 96 etc.;
4. *duration:* Duration of audio source material in time, i.e. HH:MM:SSSS format;
5. *sampling_frequency:* The rate at which the audio was sampled e.g. 44.1KHz, 96KHz, etc.;

B.4 VideoMD: Video Technical Metadata Extension Schema

VideoMD is a XML schema to describe the technical characteristics of digital video objects.

VideoMD contains eight top level elements:

1. *color:* Information describing color characteristics and specifications;
2. *compression:* The type and amount of digital compression, e.g. Predictive 10:1, RLE 2:1;
3. *data_rate:* The data rate of the video source item in Mb/s, e.g. 4.0, 8.25, 100.0, etc.;
4. *duration:* Duration of video source item in time, i.e. HH:MM:SSSS format;
5. *frames:* The number of frames and frame rate of video source item;
6. *resolution:* The horizontal and vertical dimensions in pixels and aspect ratio of the frame;
7. *sound_field:* The digital sound format used in the video source item, e.g. mono, stereo, DTS, etc.
8. *video_format:* Information describing the format specifications of the video

C. APPENDIX – Technical Metadata: Element Recommendations for Media Type Formats

As noted, for **LORs** we recommend a lightweight approach based on PREMIS with media-specific metadata extensions. In drafting our recommendations, we have also used the NZNL Preservation Metadata set to guide us in the selection of recommended elements from media-specific standards. We have provided additional information which should support the establishment of local workflows, helping to set priorities and procedures for the collection and storage of technical metadata.

The first table below represents the elements that should be collected, when applicable, for any type of content followed by media-specific tables. Each table includes the metadata elements defined by the standard listed in the table header. Highly recommended metadata are bolded, and the row is marked with grey color. All tables include extra information such as:

Could be extracted from file? This indicates whether the 'carrier' of the metadata is the file itself and whether this data can be extracted automatically using some file validation tool (Jhove, NZNL Metadata Extractor, DROID). Although the extraction procedure may be straightforward, we would still recommend storing values when possible as it eases the workflow within the local system -phased additions of data can cause synchronization problems. If digitization is outsourced, it is highly recommended that the external vendor collect whatever metadata possible and deliver it in a machine-readable form. In general, we recommend that metadata that cannot be extracted or is not carried by the file itself be recorded in some manner.

PREMIS Equivalent: For media-specific metadata standards, the PREMIS equivalent fields are listed. If an element exists in PREMIS, the HOPE recommended practice is to use the PREMIS element, instead of-or in addition to-the media-specific metadata.

Basis: For PREMIS elements, specifies whether data should be collected on masters only or derivative copies as well. As a rule, metadata collected for masters and derivatives can be similar. The only exception we have highlighted is "digital signatures", which though not currently supported by HOPE, would seem to be appropriate only for master files.

Granularity: Specifies the level at which the metadata should be collected in the case of a compound object; this can be object or file level.

OAIS Concept: Defines which OAIS function the technical metadata supports.

PREMIS Data Dictionary: Core Technical Semantic Units					
PREMIS Element	**Description**	**Can be extracted from file**	**Basis**	**Granularity**	**OAIS Concept**
1.5 objectCharacteristics	Technical properties of a file or bitstream that are applicable to all or most formats.	N/A	Master, Derivative	File	Content Information > Representation Information > **Content Data Object Description**
1.5.1 compositionLevel	An indication of whether the object is subject to one or more processes of decoding or unbundling.	No	Master, Derivative	File	Content Information > Representation Information > **Content Data Object Description**
1.5.2 fixity	Information used to verify whether an object has been altered in an undocumented or unauthorized way. (See: **Fixity**) With fixity the following sub-elements should be defined: ▪ messageDigestAlgorithm - The specific algorithm used to construct the message digest for the digital object. ▪ messageDigest - The output of the message digest algorithm.	Yes	Master, Derivative	File	Preservation Description Information > **Fixity Information**
1.5.3 size	The size in bytes of the file or bitstream stored in the repository.	Yes	Master, Derivative	File	Content Information > Representation Information > **Content Data Object Description**
1.5.4 format	Identification of the format of a file or bitstream where format is the organization of digital information according to preset specifications. With format the following sub-elements should be defined: ▪ formatDesignation - An identification of the format of the object. ▪ formatName - A designation of the format of the file or bitstream. ▪ formatVersion - The version of the format named in formatName.	Yes	Master, Derivative	File	Content Information > Representation Information > **Content Data Object Description**
1.5.5 creatingApplication	Information about the application that created the object.	No	Master, Derivative	File	Preservation Description > **Provenance Information**
1.5.6 inhibitors	Features of the object intended to inhibit access, use, or migration.	No	Master, Derivative	File	Content Information > Representation Information > **Content Data Object Description**
1.5.7 objectCharacteristicsExtension	A container to include semantic units defined outside of PREMIS.	N/A	N/A	N/A	Content Information > Representation Information > **Content Data Object Description**
1.6 originalName	The name of the object as submitted to or harvested by the repository, before any renaming by the repository.	No	Master	File	Preservation Description > **Provenance Information**
1.7 storage	Information about how and where a file is stored in the storage system.	No	Master, Derivative	File	Preservation Description > **Provenance Information**
1.7.1 contentLocation	Information needed to retrieve a file from the storage system, or to access a bitstream within a file. With contentLocation, the following sub-elements can be defined: ▪ contentLocationType ▪ contentLocationValue	No	Master, Derivative	File	Preservation Description > **Provenance Information**
1.7.2 storageMedium	The physical medium on which the object is stored (e.g., magnetic tape, hard disk, CD-ROM, DVD).	No	Master, Derivative	File	Preservation Description > **Provenance Information**
1.8 environment	Hardware/software combinations supporting use of the object.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.1 environmentCharacteristic	An assessment of the extent to which the described environment supports its purpose.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.2 environmentPurpose	The use(s) supported by the specified environment.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.3 environmentNote	Additional information about the environment.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.4 dependency	Information about a non-software component or associated file needed in order to use or render the representation or file, for example, a schema, a DTD, or an entity file declaration.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.5 software	Software required to render or use the object. With software the following sub-elements should be defined: ▪ swName - Manufacturer and title of the software application. ▪ swVersion - The version or versions of the software referenced in swName. ▪ swOtherInformation - Additional requirements or instructions related to the software referenced in swName. (If Applicable) ▪ swDependency - The name and version of any software component needed by the software referenced in swName in the context of using this object. (If Applicable)	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.6 hardware	Hardware required to render or use the object.	No	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.8.7 environmentExtension	A container to include semantic units defined outside of PREMIS.	N/A	Master, Derivative	Object, File	Content Information > Representation Information > **Environment Description**
1.9 signatureInformation	A container for PREMIS defined and externally defined digital signature information, used to authenticate the signer of an object and/or the information contained in the object. (See: **Fixity**)	No	Master	File	Preservation Description Information > **Fixity Information**
1.9.1 signature	Information needed to use a digital signature to authenticate the signer of an object and/or the information contained in the object.	No	Master	File	Preservation Description Information > **Fixity Information**
1.9.2 signatureInformationExtension	Digital signature information using semantic units defined outside of PREMIS.	No	Master	File	Preservation Description Information > **Fixity Information**

C-1. Table - Core Elements: These general element recommendations are applicable to files in every format. This chart includes higher-level elements, only specifying sub-elements for the recommended semantic units.

NISO Standard Z39.87					
NISO Element	**Description**	**PREMIS Equivalent**	**Could be extracted from file**	**Granularity**	**OAIS component**
ObjectIdentifier	A designation to uniquely identify the object	N/A	No	File	Preservation Description Information > **Reference Information**
FormatDesignation	Identifying the format of the object	format	Yes	File	Content Information > Representation Information > **Content Data Object Description**
Compression	Detailing which compression was used on the image file or digital object being described	compositionLevel	Yes	File	Content Information > Representation Information > **Content Data Object Description**
Fixity	Used to verify whether a file has changed or been altered in an undocumented or unauthorized way (checksum)	fixity	Yes	File	Preservation Description Information > **Fixity**
BasicImageCharacteristics	A container of imageWidth, imageHeight and PhotmetricInterpretation sub-containers	N/A	Yes	File	Content Information > Representation Information > **Content Data Object Description**
PhotometricInterpretation	Photometric interpretation is the information necessary to properly interpret the pixel values. If sub containers (colorSpace, colorProfile, YCbCr) do not contain any data or not applicable, the need not be recorded.	N/A	Yes	File	Content Information > Representation Information > **Content Data Object Description**
ImageColorEncoding	Contains information about all color encoding within an image. It is applicable to all images, whether the actual image is full color, greyscale or black and white.	N/A	Yes	File	Content Information > Representation Information > **Content Data Object Description**
SpecialFormatCharacteristics	Certain file formats have characteristics that are not common to other image file formats. Information which needs to be documented from these formats should be recorded in data elements in this section, grouped by format (JPEG2000, MrSID, Djvu).	N/A	Yes (with limitatons)	File	Content Information > Representation Information > **Content Data Object Description**
SourceInformation	Detailing the Source information related to the imaged subject. Comprised of sourceType, sourceID and sourceSize	N/A	No	File	Preservation Description > **Context Information**
GeneralCaptureInformation	Detailing the General Capture Information of the digital object. Comprised of dateTimeCreated, imageProducer, captureDevice	N/A	No	File	Preservation Description > **Provenance Information**
ScannerCapture	Detailing Scanner Capture specifics. Comprised of scannerManufacturer, maximumOpticalResolution, scannerSensor, scannerModel and scanningSystemSoftware. If an image created with a scanner, sub-elements should be used to sepcify the scanner settings used when the image was scanned.	N/A	No	File	Preservation Description > **Provenance Information**
DigitalCameraCapture	Detailing Digital Camera Capture. Comprised of digitalCameraManufacturer, cameraSensor, digitalCameraModel, cameraCaptureSettings	N/A	No	File	Preservation Description > **Provenance Information**
SpatialMetrics	Dealing Spatial Metrics specifics. Comprised of smaplingFrequencyPlane, samplingFrequencyUnit, xSamplingFrequency, ySamplingFrequency. While it is recognized that digital images can describe three-dimensional objects, this section deals only with the classic 2-dimensional projection of such objects as seen by the imaging device at any given instant in time.	N/A	No	File	Preservation Description > **Provenance Information**
TargetData	Identifying the information about targets used in the digitization process. Comprised of targetType, externalTarget, performanceData, targetID	N/A	No	File	Preservation Description > **Provenance Information**
ImageProcessing	Identifying the image editing or image transformation related data.	N/A	No	File	Preservation Description > **Provenance Information**
PreviousImageMetadata	A data element of technical metadata from previous generations of the image file recorded to document provenance and change history and to provide essential metadata that could be used to simulate return to original image data.	N/A	No	File	Preservation Description > **Provenance Information**

C-2. Table - Image Elements: These recommended elements are applicable to raster images in all formats. This chart includes higher-level elements only.

TextMD - Technical Metadata for Text				
TextMD Element	**Description**	**PREMIS Equivalent**	**Could be extracted from file**	**OAIS component**
encoding	Technical aspects of the text generation, whether analog-to-digital or born digital. Contains: encoding_platform, encoding_software, encoding_agent	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
character_encoding	Information regarding the encoding of characters within the file, including the standardized name of the character set, the byte order, the character size, and the line break mechanism.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
language	Language(s) used in work. Use ISO 639-2 codes, which are enumerated in the schema as valid text values.	N/A	No	Content Information > Representation Information > **Content Data Object Description**
alt_language	A language code/description for the text other than ISO 639-2. The alt_language element has a single attribute, authority, which may be used to record the source of the language code (e.g., Ethnologue).	N/A	No	Content Information > Representation Information > **Content Data Object Description**
font_script	The default font or script of the item.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
markup_basis	The metalanguage used to create the markup language, such as SGML, XML, GML, etc.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
markup_language	Markup language employed on the text (i.e., the specific schema or dtd). May be a URI for schema or dtd, but not mandatory.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
processing_note	Any general note about the processing of the file not covered elsewhere.	N/A	No	Content Information > Representation Information > **Content Data Object Description**
printRequirements	Any special requirements for printing the item.	N/A	No	Content Information > Representation Information > **Environment Description**
viewingRequirements	Any special hardware or software requirements for viewing the item.	software / hardware	No	Content Information > Representation Information > **Environment Description**
textNote	Any general note on material not covered elsewhere.	N/A	No	Content Information > **Representation Information**
pageOrder	The natural (language-specific) page turning order of the text (left-to-right for Latin-based script, right-to-left for Arabic, Hebrew, etc.) independent of how it is represented in the METS file.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
pageSequence	The arrangement of the page-level divs in the METS file. That is, does the first div contain the first page a user would naturally read based on the language-specific direction of the text (the beginning of the content) or the last page the user would naturally read (the end of the content)? Enumerated values are \'reading-order\' and \'inverse-reading-order\'.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**

C-3. Table - Text Elements: These recommended elements are applicable to born-digital text documents. This chart includes higher-level elements only.

AudioMD: Audio Technical Metadata Extension Schema				
AudioMD Element	**Description**	**PREMIS Equivalent**	**Could be extracted from file**	**OAIS component**
audio_block_size	Size of an audio block in bytes.	N/A	No	Content Information > Representation Information > **Content Data Object Description**
bits_per_sample	Number of bits per audio sample, e.g., 16, 20, 24, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
codec	Audio codec used.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
data_rate	Data rate of the audio in an MP3 or other compressed file, expressed in kbps, e.g., 64, 128, 256, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
data_rate_mode	Indicator whether the data rate is fixed or variable.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
duration	Elapsed time of the entire file, expressed using ISO 8601 syntax	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
num_channels	Number of audio channels, e.g., 1, 2, 4, 5, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
sampling_frequency	Rate at which the audio was sampled, expressed in kHz, e.g., 22, 44.1, 48, 96, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
audio_data_encoding	Structure for audio data	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
calibration_ext_int	Indicator that the calibration information is contained within the file or externally.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
calibration_location	Temporary location of the calibration file if it is external, e.g. URL.	N/A	No	Content Information > Representation Information > **Environment Description**
calibration_type	Type of calibration, e.g., the ITU test sequences known as CCITT 0.33.00 (mono) and CCITT 0.33.01 (stereo).	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
first_sample_offset	Location of the first valid sound byte in the file.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
first_valid_byte_block	Location of the first valid sound byte in the block.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
last_valid_byte_block	Location of the last valid sound byte in the block.	N/A	Yes (?)	Content Information > Representation Information > **Content Data Object Description**
note	Additional information or comments about the audio file.	N/A	No	Content Information > **Representation Information**
sound_channel_map	Information about the channel configuration, e.g., mapping the audio channel to their intended aural position/loudspeakers. The values represent parseable compound metadata using commas as separators, e.g., 1=left_front, 2=right_front, 3=center, 4=left_	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
sound_field	Indicates aural space arrangement of the sound recording, e.g., monaural, stereo, joint stereo, surround sound DTS 5.1, etc. MAVIS codes exist.	N/A	No	Content Information > Representation Information > **Content Data Object Description**
word_size	Number of bytes that comprise a single sample of audio data, which generally maps to bits_per_sample. Files with a bit depth of 24 will usually be expressed as a 3-byte word_size	N/A	No	Content Information > Representation Information > **Content Data Object Description**

C-4. Table - Audio Elements: These recommended elements are applicable to audio files. This chart lists all elements.

VideoMD: Video Technical Metadata Extension Schema				
VideoMD Element	**Description**	**PREMIS Equivalent**	**Could be extracted from file**	**OAIS component**
aspect_ratio	The desired aspect ratio of the image on screen, e.g., 4:3, etc. Some files produced for display on non-square-pixel monitors have a desired aspect ratio that differs from the ratio of horizontal to vertical pixels.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
bits_per_sample	The number of bits of sample depth, e.g., 8, 24, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
codec	Type of video codec used.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
data_rate	Data rate of the audio in an MPEG or other compressed file expressed in mbps, e.g., 8, 12, 15, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
data_rate_mode	Indicator that the data rate of the video is fixed or variable.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
duration	Elapsed time of the entire file, expressed using ISO 8601 syntax	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
frame_rate	The number of frames per second at which the video source item was digitized.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
resolution	Resolution of digital video source item expressed as horizontal lines.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
sound	Indicator of the presence of sound in the video file. If the value \"yes\" is selected, then the video file will also be associated with an instance of audioMD (audio metadata).	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
calibration_ext_int	Indicator that the calibration information is contained within the file or externally.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
calibration_location	Temporary location of the calibration file if it is external e.g. URL	N/A	No	Content Information > Representation Information > **Environment Description**
calibration_type	Type of calibration used.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
note	Additional information or comments about the video file.	N/A	No	Content Information > **Representation Information**
pixels_horizontal	The horizontal size of a frame in picture elements.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
pixels_vertical	The vertical size of a frame in picture elements.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
sampling	The video sampling format (in terms of luminance and chrominance), e.g., 4:2:0, 4:2:2, 2:4:4, etc.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**
scan	Indication whether digital video item is scanned in an interlaced or progressive mode.	N/A	Yes	Content Information > Representation Information > **Content Data Object Description**

C-5. Table - Video Elements: These recommended elements are applicable to video files. This chart lists all elements.