Deliverable

HOPE

Grant agreement no: 250549

Heritage of the People's Europe

HOPE: Mission, Technical Vision and High-Level Design of the Architecture

Deliverable number	D 2.1
Status	Final
Authors	Sjoerd Siebinga
	Paolo Manghi
	Mario Mieldijk
	Titia van der Werf
Delivery Date	31 August 2010
Revisions	24 June, 30 August 2012
Dissemination level	Public

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

Version history

Date	Changes	Version	Name
31.08.2010	First version of the document	1.0	Sjoerd Siebinga Paolo Manghi Mario Mieldijk Titia van der Werf
24.06.2012	Clarified the end user position throughout the document; added TOC; added Introduction to better explain Vision, HLD – Architecture and their relationship; replaced Glossary with a new version 2.0	1.5	Repke de Vries
31.08.2012	Added Mission (what HOPE is about); made Vision explicit as first of all Technical Vision (what HOPE tries to achieve IT-wise and how: BPN); added cross-references to the HOPE Glossary to better explain domain and other specific terminology; added archival terms to Glossary (now V2.1)	2.0	Erik-Jan Zürcher Cristiana Pipitone Repke de Vries

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

General introduction: the field and mission of HOPE

Throughout history, documentary records have generally been generated and kept by political and religious elites, primarily courts and states. In doing so, these elites have shaped our picture of history. The voice of the peoples of Europe therefore is heard primarily in an indirect fashion: filtered by the records of those in power, in which they figure as tax payers, recruits, criminals and litigants, but are rarely accorded agency.

It was the industrial revolution that changed this state of affairs and it did so in two separate, but interrelated, ways. From the mid-Nineteenth Century onwards a growing interests developed in the living conditions of the new industrial proletariat. People like Friedrich Engels in Britain and Frederic le Play in France investigated conditions in the factories and slums. By the end of the Nineteenth Century this interest led to the foundation of the earliest museums on industrial labour, all of them focused on issues of health and safety. The most famous of the early institutions was the Musée Social founded in Paris in 1894, which also included an important library on social issues and which served as the model for similar institutions in Europe and America. The aim of these institutions was not to preserve the history of the working class – it was to instruct people and improve industrial practices.

The second development that gave a voice to working people was the emergence of mass organisations, trade unions and parties. These organisations kept records themselves, but conditions for preserving their documentary heritage were inauspicious. The workers organisations were poor, certainly in their early years, and therefore used cheap paper and ink. More importantly: they were often persecuted and suppressed. So many records were lost (only one page of the manuscript Communist Manifesto survives, for instance). Nevertheless, movements that were in a position to do so, such as the Swedish trade unions, founded institutions to keep their archives and libraries around the same time as the first social museums – in the 1890s. In 1921 D. Riazanov started the Marx-Engels Institute in Moscow,that brought together a very important collection on early socialism hot on the heels of the Russian revolution.

The interbellum and World War II brought new dangers to the heritage of social movements. Financial difficulties during the first world crisis and factionalism within the movements led to parts of their documentation ending up in different places, but above everything else it was the dictatorships of the Nineteen Thirties, in Italy, Spain, the Soviet Union and Germany that endangered the collections. The International Institute of Social History in Amsterdam was founded in 1935 with the express purpose of offering a place of refuge for these collections. Under German occupation many of the collections that had been built in the last half-century were confiscated and removed by the Nazis, as they represented the heritage of their arch enemies: anarchists, communists, socialists, trade unionist and Jews.

In the new democratic environment in post-war western Europe the old collections were rebuilt and many new institutions founded and in 1970, when interest in social movements was at its peak following “the events of May 1968”, fourteen partners founded the International Association of Labour History Institutions (IALHI) to further cooperation and exchange between the members. In the past forty years IALHI has grown into a network of around eighty partners, some of them large and well-

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

supported and some of them tiny. They include archives, libraries, universities and museums all over Europe and some beyond Europe's borders.

The HOPE project, in which twelve leading IALHI members take part, makes it possible to create a shared platform with state of the art technology and procedures and thus to achieve four things for the field:

1. To create enhanced visibility for the IALHI collections, both among the public at large and among specialists;
2. To increase cooperation and exchange between IALHI members (the original IALHI aim);
3. To allow smaller institutions with important material but few resources to profit from the infrastructures of stronger sisters, and
4. To reconstruct virtually collections that have become scattered as the result of strife, revolution and war in Twentieth Century Europe.

HOPE will make a major contribution to ensuring that the authentic voice of the working people of Europe, as preserved in the private collections of IALHI rather than in the records of the state, continues to be heard.

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

HOPE Technical Vision

INTRODUCTION

To build the shared platform or *HOPE System* [G(lossary) (Sub Section) 1] with state of the art technology and procedures mentioned earlier, first a Technical Vision has to be developed and based on that the High Level Design of the architecture for the platform.

This deliverable addresses both: Technical Vision and High Level Design.

Vision formulation is based on the OpenUp methodology[1] with these main elements:

- Stakeholder descriptions
- Describing the environment of the future users of HOPE
- Outlining the core requirements and constraints from different perspectives: those of the targeted users and those of the content providers

Given the software development roots of OpenUp, language and logic of the above formulation are technically oriented but explicitly cover the so called "[end] user dimension" as well.

In line with the OpenUp philosophy, the vision provides a strategy against which all future technical decisions can be validated. It also rallies the different partners in the project around the technical end goals and gives them the context for decision-making in their respective Work Package tasks. Next to that the vision is input for the High Level Design (HLD) of the HOPE Architecture.

The High Level Design identifies the basic concepts, principles, functions, data flows and open standards of the HOPE System as shared, web-based *discovery-to-delivery* [G 1] service. It thereby serves as starting point and baseline for all the HOPE System detailed specifications and deliverables planned later in the project as Work Packages 1, 2, 3, 4 and 5. From these Work Packages, user needs are the concern of WP1 "Users, Content and IPR".

A series of diagrams visualises the High Level Design and are available in Appendix.

Also the HOPE Glossary can be found in Appendix. It is ongoing work providing definitions of acronyms and abbreviations and a terminology. Throughout the text presenting key terms *in italics* together with Glossary sub section indication (like: [G 1] gives the reference to this Glossary.

[1] http://epf.eclipse.org/wikis/openup/

POSITIONING

Objectives

The general introduction to this deliverable concluded with four achievements for the field. Each is repeated here and translated to a concrete objective for the Technical Vision of HOPE.

The first and second one:

- To create enhanced visibility for the IALHI collections, both among the public at large and among specialists
- To increase cooperation and exchange between IALHI members (the original IALHI aim)

Project HOPE aims to improve access to the vast amount of highly significant but scattered digital collections on social history across Europe. It proposes to achieve this through a "best practice network" that will bring about greater collaboration among the libraries, archives and museums of the IALHI members, who share the same purpose, which is to advance scientific and general knowledge of social history.

The short-term objective of the *HOPE best practice network* [G 1] (BPN) is to implement the HOPE System, that will make the digital *collections* [G 1] of the participants available through Europeana[2] and other *discovery services* [G 1]. The HOPE System consists of the local systems of *Content Providers* [G 1] (CP), the *HOPE Aggregator* [G 1], the *HOPE PID service* [G 1, 3], the *HOPE shared object repositories* [G 1] and the *discovery services* [G 1].

The third and fourth one:

- To allow smaller institutions with important material but few resources to profit from the infrastructures of stronger sisters, and
- To reconstruct virtually collections that have become scattered as the result of strife, revolution and war in Twentieth Century Europe.

The longer-term objective of the HOPE best practice network is to share data, services and expertise and in doing so to achieve economies and efficiencies that permit the (digital) collections in the libraries, archives and museums to be effectively described, comprehensively disclosed, successfully discovered and appropriately delivered. The BPN is therefore geared from the start towards the adoption of best practices by the content providers and not towards implementing centralized ICT solutions which will be customized to the particularities of individual CPs, thereby locking them in the HOPE system.

[2] http://www.europeana.eu

Problem Statement

The lack of technical expertise, interoperability and shared best practices, resulting in disconnected catalogues and fragmented access to digital collections, affects the HOPE content providers and IALHI members, the impact of which is:

- Hampering the research process and research productivity in the social and historical sciences
- Severely hampering the discovery experience of the general public

A successful solution would be:

- Harmonization and adoption of best practices in digitization, use of *metadata* [G 2], presentation of collections, content delivery, etc...
- Integrated access to the shared collections on social history and the different material types, previously only available through individual, dedicated systems
- Bringing the collections to the users in a comprehensive way by populating discovery services (such as Europeana) with metadata about the collections
- The unambiguous and seamless navigation from a search result in any given discovery service to the digital resource (d2d logistic) (*HOPE Social History Resource* [G 1]*)*.

Position Statement of the HOPE System

The HOPE System as shared platform is a web discovery-to-delivery service infrastructure that will serve a variety of users, ranging from the scientific researcher to the general public, who wants to find, access and use social history resources (see also: *Archival Finding Aid* [G 2]).

It provides uniform web-access to large amounts of historical materials of very rich diversity.
Unlike any other European social history resource published online to date, the HOPE System is most encompassing in terms of content, scale and accessibility.

STAKEHOLDER DESCRIPTIONS

Stakeholder Summary

In project HOPE there are five different types of stakeholders. For each category a brief description and summary of responsibilities is given below.

Given the fact that “Content Providers” are identified as Stakeholder and given the fact that HOPE aims to bring “content” and “users” together in new, innovative ways that should be rewarding to both, one could argue the need to include the targeted HOPE user groups as Stakeholder. There is no organisational structure though to have users formally represented as Stakeholder. Instead the project has a Work Package that concerns it self with user needs.

Please also see “User Environment” right after “Stakeholder Summary” – both under the same header “Stakeholder Descriptions”.

1. Content providers

These are organisational-units (library, archive, museum) managing social history collections. In HOPE the following eleven organisations (all leading IALHI members) are content providers: KNAW-IISG, AMSAB-ISG, CGIL, FES, FMS, SSA, TA, VGA, KEE-OSA, UPIP and GENERI.
The HOPE project website www.peoplesheritage.eu has further pointers to these organisations.

Their responsibilities are:

- To provide the content (metadata, links to the *digital objects* [G 4] and/or files of the digital objects)
- To ensure the agreed level of quality of the metadata
- To carry out the local implementations required for the supply of the content to the HOPE aggregator and HOPE repository or to implement a local *HOPE-compliant repository* [G 1] if the partner is not making use of the HOPE repository.
- To implement the necessary workflows for the supply of content in the future (also after the lifetime of the project)

2. Technology providers

These are organizations with technical expertise, capacity and facilities, operating parts of the HOPE system. The following organizations are technology providers: KNAW-IISG and CNR-ISTI (also see the HOPE project website).
Their responsibilities are:

- To implement the HOPE repository system and the HOPE PID service (KNAW-IISG)
- To implement the HOPE aggregator system (CNR-ISTI)
- To ensure the operation of the system at the agreed service level

3. Web discovery service providers

These are organizations operating a discovery service on the Web and participating in the HOPE project: EDLF and KNAW-IISG.
Their responsibilities are:

- To operate Europeana (EDLF)
- To operate the *IALHI Portal* [G 1] (KNAW-IISG)

4. Coordinator

Organization coordinating the HOPE project: KNAW-IISG.
Its responsibilities are:

- To ensure the HOPE Description of Work (DoW) is carried out
- To monitor and report on the project's progress

5. Funder

Funding Organization of the HOPE project: The European Commission (EU)
Its responsibilities are:

- To monitor the project's progress
- To approve funding

User Environment

The user environment is a given web discovery service or *social site* [G 1] on the web, where the user usually goes to for searching or browsing cultural heritage collections or social history resources. This may be any service containing metadata (and *previews* [G 4]). It could be Europeana, Gallica, the IALHI Portal, the institutional website of the FMS, the photo stream of the IISG on Flickr, WorldCat, Google, or any other site. In this heterogeneous environment the user will find HOPE metadata, previews of the corresponding *digital objects* [G 4] and a link to the *digital* object *files* [G 4].

How the user finds this metadata and the discovery process itself is mostly out-of-scope: the HOPE system interfaces in various ways with discovery services but it usually cannot influence the behaviour of the search or discovery system or social site. The only exceptions are the discovery sites that use the Search web service (*HOPE Search API* [G 1]) of the HOPE Aggregator (the IALHI Portal and the HOPE partner websites).

From the moment that the user finds HOPE metadata of specific collection *items* [G 2] of his/her interest, the way in which the user is led to the actual digital resource and experiences each step in the *d2d* [G 1] process (request, locate, retrieve, access, consult via a reader/player, download or request a copy in a higher resolution or a print reproduction, online payment, contact with the service desk, etc.) is critical to user satisfaction.

Once the d2d process has resulted in a delivery (e.g. download of the requested files), the items of interest can be used in the user's content processing environment (Photo-editing, text-processing, text-mining, content mash up, etc.). The user's content processing environment is out-of-scope, but links referencing to the source of the content (metadata and digital objects) should be provided and the links should unambiguously refer back to the original content.

HOPE SYSTEM OVERVIEW

Needs and Features (content providers and target users perspective)

Below a list of capabilities needed, from different perspectives (content providers, target users) and why they should be implemented by the HOPE system, with the corresponding feature description.

1. Content providers want to populate web discovery services with their metadata in order to maximize the chances that users find their collections => need for an aggregator that gathers all the metadata within a given domain (social history) and that disseminates the metadata to the different discovery services.

2. The user is highly dependent on the quality of the metadata to be able to ascertain if a specific digital object is of interest => need for effective discovery metadata with an unambiguous reference to the (trusted) source (see below point 19)

3. The user is dependent on the representativeness of the preview to be able to ascertain if the corresponding digital object is of interest => need for representative previews for all material types.

4. The user expects predictable search functionality => need for a Search web service within the HOPE system (part of the Aggregator function) with standard functionality

5. The user expects to find "all" relevant hits during a search => need for maximal support of multilinguality by the Search web service within the HOPE system and maximal semantic interoperability at the metadata level

6. The user expects to locate, retrieve and access the digital object via a stable and persistent link on the web => need for unique *persistent identifiers* [G 3] and *resolving* [G 3] mechanism.

7. The user needs to consult the *digital object* [G 4] via a reader/player => need for support of common file formats of readers/viewers

8. The user needs to download a digital copy or a *derivative* [G 4] of the digital object, for reuse in another environment => needs for support of conversions to a wide variety of formats

9. The user wants a print of the digital object in original format (e.g. poster) => Printing service organization that delivers print reproductions to a customer, based on digital objects from the HOPE compliant repository => need for support of delivery requests from third parties (e.g. reproduction service; web-shop)

10. The user needs to ask a question or report a problem during the delivery transaction => need for Helpdesk functions and staffing to give support to users

11. The content provider needs to set policies and to manage its own content, which he has submitted to the Aggregator and to the HOPE repository => need for content provider admin dashboards.

OTHER REQUIREMENTS AND CONSTRAINTS

Below are listed, at a high level, platform requirements, performance requirements, and environmental requirements; design constraints; external constraints, assumptions or other dependencies.

12. The HOPE aggregator function will be an implementation of the existing D-Net platform and will therefore be defined by D-Net's capabilities and constraints. D-Net is an open source platform developed by CNR-ISTI.

13. The HOPE repository function will be an implementation of existing open source solutions, selected to fit HOPE's needs as closely as possible, and it will therefore be defined by the chosen software's capabilities and constraints.

14. More generally, the HOPE system will adhere to open architecture principles and make use of open standards and open source solutions. The choice for open source solutions permits more (cost-)effective investment in shared development (requirements management, code development and testing) and in the building of shared technological know-how within the HOPE community, and minimizing dependency on third parties with vendor lock-in implications. Opting for open source also brings its own constraints and risks, such as lack of support, steep learning curve, extended response time to troubleshooting, etc. The HOPE BPN

however chooses to be a BPN not only in its own domain but also in the underlying domain of software solutions.

15. The HOPE aggregator services (incl. the Search web service), the HOPE PID service, and the HOPE compliant repositories (incl. the SOR) need to be available 24/7.

16. The content providers should adopt the agreed best practices and not become dependent of the HOPE aggregator function for their metadata production process. They should keep the capability of producing harmonized HOPE metadata, also without the aggregator function.

17. The content providers should adopt the use of persistent identifiers for their digital objects and not become dependent of the HOPE content repository for the storage of their digital objects. They should be able to transfer their content to another repository without breaking the access link to the objects.

18. The content providers should adopt the use of persistent identifiers for their metadata records and *authorities* [G 2] and not become dependent of the information space of the HOPE aggregator or of Europeana, or any other sub-space of the web. Any metadata record and authority record populating a web discovery service should always refer back to its original and trusted source: the record maintained by the content provider. This is necessary for guaranteed access and verification and authentication purposes.

19. The content providers should be encouraged to locally reintegrate the delivered metadata after they have been enriched and harmonized by the HOPE Aggregator. If this is not feasible, the content providers should be encouraged to enrich and harmonize their metadata according to the HOPE best practices as much as possible before delivering them to the Aggregator. Metadata kept locally and disseminated across web discovery services should be synchronized as much as possible

20. Europeana promotes public access and unrestricted access to the underlying content and stimulates content providers to clear the *IPR* [G 5] of their collections. HOPE has identified within its collections the content with public access in Table 0 of the DoW. However, HOPE also acknowledges that its partners hold collections with *access and use restrictions* [G 5]. Even if Open Access principles are to be pursued actively, privacy issues and *copyrights* [G 5] have to be respected. That is the duty of archival institutions and determines their trustworthiness. Therefore the HOPE system needs to support and facilitate access and use restrictions.

21. The content providers should maintain information on access and use restrictions in the repository, as these restrictions concern the digital objects stored in the

repository. These metadata need not be integrated in the descriptive metadata disseminated by the Aggregator, as they are pertinent to delivery only, not to discovery.

PURPOSE AND SCOPE

The High Level Design (HLD) of the HOPE Architecture identifies the basic concepts, principles, functions, data flows and open standards of the HOPE system. The design is visualised in the HLD Diagrams, which are available in Appendix.
In the coming sections and as compendium to these diagrams, the decisions taken for the high-level design of the HOPE System and the workflow of the diagrams will be explained. As will the architecturally significant requirements, constraints, decisions and objectives be indentified and analysed. The sections have a strong technical and design focus, but on a high level.
We have tried to find the right balance between HLD and describing the technical consequences of the choices made. However, we must stress that the following is neither a Functional Requirements nor a Technical Specifications document. It is not providing any comprehensive or detailed listing of the data requirements, conversion specifications, data flows, interfacing requirements, database requirements, etc. Rather it will serve as a starting point and baseline for all the detailed specifications and deliverables planned later in the project (WP1, WP2, WP3, WP4 and WP5).

Incremental approach and Roadmap for implementation

Also the high-level design approach conforms to the OpenUp practices, which we intend to follow during the design, development and implementation of the HOPE system. See for more information, the OpenUp wiki: Evolutionary Architecture[3].

The first version of the HOPE high-level design focuses on outlining the initial architectural decisions and forms the baseline on which HOPE Work Package 2 and the other Work Package Tasks can start to build.
Work on the high-level design will continue:

- By gathering input from the consensus building and best practices work packages (WP1 [also covering user needs] and WP2) and from the detailed implementation designs from Work Packages 3, 4 and 5
- By guiding the detailed implementation designs in order to ensure conformance to the high-level architectural decisions taken.

New iterations of the HLD-documents will be issued when necessary. Further refinement of the component parts of the architecture will be documented in separate documents, as part of the effort of Work Packages 4 (HOPE Aggregator) and 5 (HOPE Shared Object Repository).
All unresolved design and implementation issues are listed in the HOPE Design & Implementation Roadmap, which is an (internal) output of the high-level design task.

THE SIX MAIN SECTIONS

Against the background of the HOPE Vision and under reference to the Glossary, the following are the six main sections:

[3] Evolutionary Architecture

- The discovery-to-delivery process
- The HOPE system and high-level workflow
- The systems of content providers interfacing with HOPE
- The HOPE Persistent Identifier (PID) Service
- The HOPE Aggregator
- The HOPE Shared Object Repository (SOR).

Section 1 identifies the major overall requirements for the HOPE system to realize the unambiguous and seamless navigation from a search result in any given discovery service to the digital resource (d2d logistic).
Section 2 provides a high-level overview of the major functional components of the system and of the intended behaviour of the system to be captured in use cases.
Section 3 identifies the local systems of content providers (CP), which are required to interface with the HOPE system and describes in some detail how the metadata and content of CPs can be integrated in the HOPE system.
Section 4 describes the high-level functionality of the HOPE Persistent Identifier Service and the main considerations why *PID*s [G 3] play such a critical role in the HOPE data-flow.
Section 5 describes the high-level functionality of the HOPE Aggregator, based on the existing D-NET system and highlights specific HOPE requirements to be implemented in D-NET.
Section 6 describes the high-level functionality of the HOPE Shared Object Repository and the main design considerations, components and API data-flows.

1. Discovery-to-delivery proces

As stated in the Vision, the HOPE System is a web discovery-to-delivery service infrastructure that will serve a variety of users, ranging from the scientific researcher to the general public, who want to find, access and use social history resources.
The journey between discovery and delivery of information resources on the Internet is accomplished with a variety of differing technologies and processes, many of which fall under the responsibility of different providers (discovery services, aggregators, repositories, etc.). In HOPE we need to take decisions concerning choices to improve d2d.

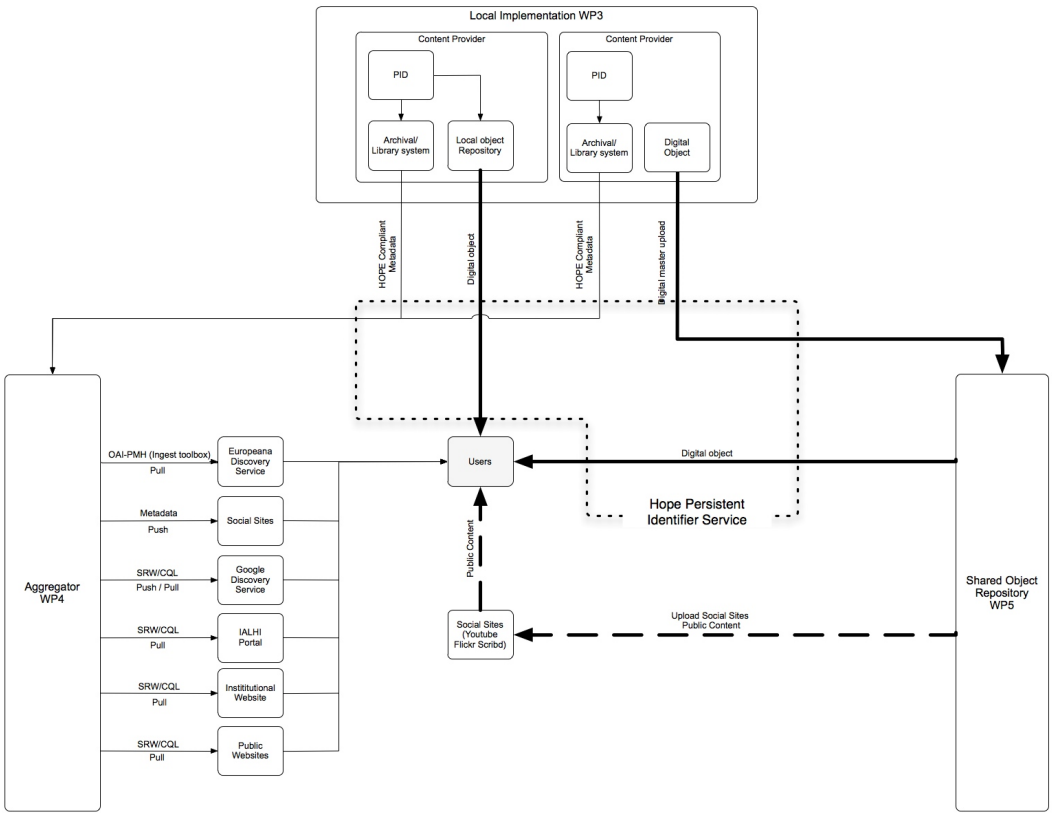
Important ways to ensure a seamless d2d process and to meet user needs and expectations are the use of:

1. *Open architecture principles for flexibility of infrastructure and interoperability of technical solutions;* the use of open standards ensures that there are as few dependencies as possible between various software components working together, and between different information systems that communicate with each other. The starting point for HOPE is to make maximal use of web-standards and standards for the exchange of data within the heritage sector and the research community (eg. SRU and OAI) - *furthermore* by limiting the choice of technological components to a number of proven open standards and open source solutions, it becomes possible to invest effectively in technological expertise and knowledge development within the HOPE community, minimizing dependency on third parties with vendor lock-in implications.

2. *Best Practices for predictability and quality improvement of the d2d process;* there is inherently much variety in customer demand, hence the need for flexible service delivery to absorb that variety - but at the same time users also expect predictability in the way services are delivered, predictability in the way search services and delivery trajectories are offered - hence the need for harmonization of practices and adoption of best practices.
3. *Internet protocol* (TCP/IP), in particular HTTP across all applications;
4. *Using URIs to uniquely identify information resources (digital objects and metadata) across the Internet* and beyond specific applications and information domains (e.g. loc-nr in WorldCat). In HOPE we adopt PIDs (i.e. persistent resolvable URIs) for all our digital objects and descriptive metadata units.
5. U*sing simple web API's to integrate with the HOPE components and other web-services.* HOPE learns from the best practices of the programmable web. Interaction with HOPE should be transparent for both the general web-user and application developers. Being part of the web is one of key design criteria for HOPE.
6. *Easy retrievability and use of public content,* without bothering users with online statements for the only purpose to protect CPs from any liability risk. The IPR guidelines (WP1) will work out these principles in more detail.
7. *Improving single sign-on authentication*: in HOPE we will not support single sign-on across d2d platforms. The starting point is that HOPE is an open, www-wide information space – it does not support closed systems. The assumption is that **any** web user may have found a link to an object file from the HOPE social history resource from **any** discovery service. The HOPE compliant repository is discovery context-agnostic. The repository serves any web user requesting a digital object file, based on its Identification Authentication and Authorisation (IAA)-system. Community members of HOPE Partners can access the objects via API keys provided to the HOPE partners. The use of API keys ensures that if they want to display restricted objects on their local websites they can do this without additional login steps or cumbersome merging of login information.

2. HOPE system and high-level HOPE data-flows

The HOPE system consists of the local systems of Content Providers, the HOPE Aggregator, the HOPE PID service, the HOPE SOR and the discovery services. Below a diagrammatic representation of the component parts of the HOPE system and of the data-flows can be found.



This section describes briefly the key components and identifies a number of key actors, the main use cases, and supporting requirements for the HOPE system as a whole. The basic flow of each of the use cases is outlined briefly and not described in any detail. This needs to be done in next iterations and the resulting use cases will be documented in other documents. However, it is important to note that in the HOPE data-flow the digital masters must be submitted to the SOR before the descriptive metadata can be submitted to the Aggregator.

The HOPE system consists of 3 core components:

1. The HOPE PID Service: which provides resolvable PIDs for the HOPE digital objects and metadata records and thereby ensures the d2d process;
2. The HOPE Aggregator: which integrates and harmonises all the metadata records of the HOPE CPs and disseminates the metadata to web discovery services;
3. The HOPE SOR: which keeps the digital masters of collections held by the HOPE CPs and delivers copies and derivates on demand;

These core components interact mainly with 4 types of actors in the environment surroundings:

1. Local systems of Content Providers: systems with which the CPs produce/manage their metadata and their digital collections;

2. Discovery Services: which provide search and browse facilities to discover the integrated content from Aggregators and other information resources on the web.
3. The web user: who makes use of the Discovery Services and of the delivery services of the SOR.
4. Social sites: which facilitate the sharing of digital content with/between web users and are generally focused on one type of content medium (text, video, pictures, music, etc.) or networking purpose (research, swimming, blogging, gaming, etc).

Main use cases:

1. CP supply-use-cases:
    a. Digital master upload by the CP to the SOR: The digital objects supplied to the SOR may come from any local network-accessible system used to store and access digital objects (eg. a digital assets management system, an FTP/HTTP-server or even a HOPE compliant repository). If the CP stores digital objects on offline carriers, he can choose to set-up a local network-accessible system for the supply of digital objects or to upload digital objects directly to the staging area of the SOR. The digital objects supplied should conform to the agreed HOPE standards, in particular they should each come with a PID.
    b. Supply/Harvesting of HOPE compliant metadata from the CP to the Aggregator: A system is required for assembling the exports from the Archival/Library system (the bibliographic records and archival finding aids) into metadata batches (data-sets), ready to be collected by the HOPE Aggregator. This system might be an OAI-PMH repository or an FTP-server for example. The metadata should be HOPE compliant, in particular they should contain the PID of the descriptive record and the PID of the corresponding digital object. This PID may be allocated by the local PID Service or fetched from the HOPE PID Service.
2. Aggregator export-use-cases:
    a. Metadata export from the Aggregator to Europeana: this export conforms to the Europeana Metadata Schema and to the export protocol supported by Europeana (OAI-PMH harvest and ingest into the Europeana Ingest Toolbox).
    b. Metadata export from the Aggregator to Google: this flow ensures that the HOPE metadata are supplied to the Google crawlers.
    c. Metadata export from the Aggregator to the IALHI Portal: this is the API interface of the Aggregator Search webservice.
    d. Metadata export from the Aggregator to the CP institutional websites: this is the API interface of the Aggregator Search webservice.
    e. Metadata export from the Aggregator to Public websites: this is the API interface of the Aggregator Search webservice.
3. User request-use-cases:
    a. User requests a digital object from the local HOPE compliant repository: a web-user who has found a relevant description of a digital object follows the digital object's PID link which activates a chain of HTTP-requests via the PID-resolver mechanism of the CP's local PID

Service. The local HOPE compliant repository reacts to the HTTP-request by providing the jump-off page with links to different versions of the digital object (a choice of derivatives) and/or with transactional information for accessing the resource (rights/payment transactions).

b. User requests a digital object from the SOR: a web-user who has found a relevant description of a digital object follows the digital object's PID link which activates a chain of HTTP-requests via the PID-resolver mechanism of the HOPE PID Service or the CP's local PID Service. The HOPE SOR reacts to the HTTP-requests by providing the jump-off page with links to different versions of the digital object (a choice of derivatives) and/or with transactional information for accessing the resource (rights/payment transactions).
c. User requests public HOPE content from the social site: a web-user who has found a relevant link to a digital object from the public HOPE content on social sites, follows the link which activates the social sites rendering services (eg. a video player), through which the digital object is displayed.

4. Upload of public content to social sites: this is the flow of well-defined sets of public content which the HOPE SOR uploads to a given social site (for which HOPE has opened an account). The returned embed-URLs are added to the jump-off page information and to the CPs. Additional descriptive metadata for the description is acquired from the Aggregator Search web service and is pushed to the social sites together with the digital objects.

3. The systems of content providers interfacing with HOPE and data-supply flows

Section 2 permits us to identify the highest value and highest risk items so that we can concentrate on these first. During previous iterations of the High Level Design it became apparent that the design made some assumptions about the prerequisites for supply of data by CPs to the Aggregator (WP4) and to the Shared Object Repository (WP5). For WP3 and the first Best Practice Network Workshop to be held beginning of September 2010, it is very important to clarify up-front and as soon as possible the requirements for the Content Providers.
Our working assumption is that there are a set of minimal requirements which all CPs can fulfil, namely that the CPs:

1. Have links between their metadata records and their digital objects,
2. Are able to export their digital objects to some kind of file-system before uploading them to the SOR
3. Are able to export their metadata in XML (encoded in UTF-8) and assemble the metadata in data-sets on some kind of file-system for the Aggregator.

In addition, our starting point is that CPs are joining the HOPE BPN with the intention to adhere to the HOPE best practices (see also Vision, Objectives) - namely in terms of:

1. Providing HOPE compliant metadata,
2. Applying persistent identification to their objects and metadata records
3. Following agreed HOPE supply-procedures and work-flows
4. Supporting agreed protocols to interface with the HOPE system.

CP data-supply flows

This sub-section explains the CP supply-use-cases in some detail, highlighting the requirements. The metadata/content requirements (e.g. HOPE Metadata Schema, Content formats) need to be worked out in detail by WP2 and the use cases need to be worked out for each CP separately in the framework of WP3, but this section gives an illustration of the typical use case, with the various steps necessary to supply the descriptive metadata and the digital objects of a data-set in the HOPE system.

- Preparation by the content provider before submission of a data-set:
  - Add a persistent identifier to every descriptive metadata record. The PID must be coded in a (additional) metadata field of the metadata record in the Content Provider's own metadata management system (e.g. archival or library information system). This identifier will be used by the Aggregator to identify, store and update the metadata record.
  - Add a persistent identifier for each digital (master) file (e.g. scan of a book page) that is part of the digital object (e.g. the book) corresponding to the relevant metadata record. One digital object (ex. a book or a letter) could consist of very many digital files. Each file should receive its own PID, but there is only one description for one object. So in the description of the object you want has only one PID, that of the structural metadata-file (e.g. a METS document) in which the different scans are structured for viewing. The persistent identifier of the container will be used by the SOR to ingest/store/update the digital object and to disseminate derivatives. During ingestion, the SOR will update the resolvable links for the PIDs.
  - The content provider must be able to export the digital masters from his native store as files and submit the files paired with the persistent identifier.
  - Define a data-set: a set of descriptive metadata records that the CP wants to treat as a submission unit for the Aggregator.
  - The content provider must be able to export all descriptive metadata records as valid XML records and encoded in UTF-8. Preferably in a single file per data-set. Large files must be compressed (e.g. gzipped) to reduce network overhead.
  - The content provider should make the descriptive metadata available to the aggregator via OAI-PMH or FTP.
  - The content provider is responsible for pushing his digital objects to the Shared Object Repository first and then to allow the aggregator to pull the corresponding data-set with metadata records. The pairing is done on the metadata level. The CP is responsible for including the PID of the digital master or compound object in each metadata record provided to the Aggregator.
- Submission of digital masters to the Shared object repository
  - Initial submission
    - CP: Content Provider creates a XML processing instruction file that contains all the necessary information to make the submission API call to the SOR. This information includes amongst others: the persistent identifier, mime-type, access

information, temporary location of the digital object for ingest (this can be on local disk, URL, or as part of the HTTP post), API-access key, content checksum of the digital object
        - CP: The Content Provider uses the SOR submission tool to process XML processing instruction and submits the digital objects to the SOR.
        - SOR: When the object is ingested the SOR ingest platform will update the resolve URL for the persistent identifier in the respective persistent identifier service to resolve to the SOR dissemination API.
    - Re-submission (updating)
        - Updating uses the same mechanism as the initial submission. Only the content checksum will be used to check if the same digital object is already stored. In that case only the technical metadata is updated. When the content hash is different the old digital master is disconnected from that PID and deleted providing no other PIDs are connected to it.
    - Deleting digital masters
        - For deleting digital masters again use the XML processing instruction. The API call then issues a delete request and changes the resolvable URL associated with the persistent identifier that is deleted to no longer point to the SOR.
- Submission of descriptive metadata to Aggregator
    - Initial submission
        - CP: Register as Provider with the Aggregator
        - CP: Register the data-set with the Aggregator
            - this includes OAI-PMH or FTP information so the Aggregator can harvest the descriptive metadata
        - CP: Create the metadata mapping following the BPN guidelines.
        - CP: Submit the mapping file to the Aggregator
        - Aggregator: The maintainers of the Aggregator software are responsible for transferring the mapping rules into the Aggregator.
        - CP: Flagging the data-set for first ingestion run in the Aggregator management interface
        - Aggregator: The Aggregator harvests the data-set and inserts the metadata records with the persistent identifiers.

      After the initial harvest, all records are available in the Aggregator for exposure to discovery services, in the metadata formats supported by the Aggregator
    - Re-submission (updating)
        - The Aggregator will periodically check for updates and re-ingest the data-set. New records are inserted, modified records are updated, and deleted records are flagged as deleted. For purposes of re-aggregation (e.g. in Europeana) the PID of the deleted records is never deleted from the system, but just flagged as deleted. When at a later stage it is resubmitted, it will be flagged as active again.

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

        - ■During re-submission the data-set will not be locked, but the latest version is still available for re-aggregation/exposure.
    - Deleting metadata records
        - ■When either the OAI-PMH protocol is not properly implemented (i.e. without persistent deletions) or the metadata is delivered via FTP, a mechanism must be defined in the Aggregator to handle deleted records. For example, all records that were not modified after a re-submission will be flagged as deleted.
- Dissemination:
    - The Aggregator is able to disseminate the metadata records in a number of different metadata formats and different methods of access to the metadata - such as OAI-PMH, SRW, and plain export - both on record and data-set level.
    - All access to the object stored in the Shared Object Repository will go via the PID of the digital master that is supplied during ingestion.

HOPE compliant repositories

CPs can choose to make their digital content available through a local repository system, but in order to fit in the HOPE system, this local system needs to comply to a minimum set of requirements. The local system may be a a local instance of the SOR, in which case it will be HOPE compliant, but it could also be a newly acquired repository, or an enhanced version of the local systems. In such cases, the requirements for HOPE compliancy need to be made explicit. This will be worked out in more detail in the next iteration of the HLD (see Design & Implementation Roadmap).

4. HOPE Persistent Identifier Service

The HOPE BPN has chosen to require Persistent Identifiers for all metadata records and digital objects submitted to the HOPE system (see Vision). The usage of these PIDs is very important to the maintainability of the HOPE system. Duplicates' detection and merging has been one of the core problems of large-scale aggregation and unification of heterogeneous sources. The consistent implementation of persistent identifiers by Content Providers will significantly simplify the work-flow and enhance the reliability of the HOPE system.

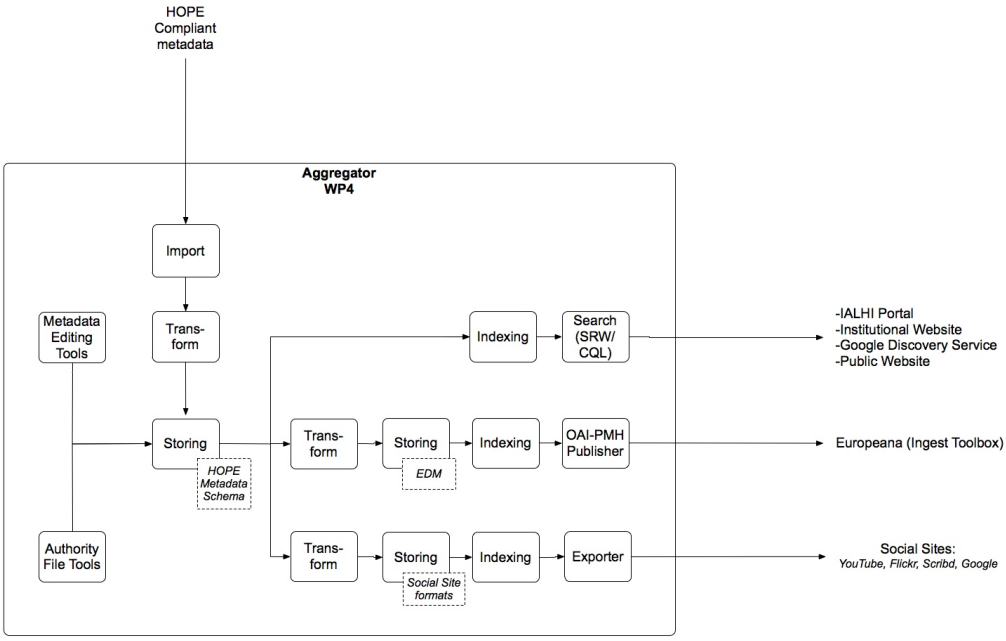
The PIDs submitted to HOPE must be registered at a known persistent identification web-service, such as HANDLE, PURL, DOI, ARK, etc.. WP2 will provide recommendations concerning which PID services conform to the HOPE requirements. The addition of persistent identifiers as resolvable URLs to digital objects will also give Content Providers much flexibility and control over the resolvability of the objects they refer to. For example, when the CP decides to migrate the digital objects to another repository they will only have to change the resolve URL that the PID refers to in the respective web-services. It will not be necessary anymore to issue new URLs for all these objects to all the users and services that are using or have bookmarked these objects.

The HOPE system will also provide its own Persistent Identifier Service, because not every Content Provider is able to host its own system. Since there are many PID services available on the web today, the HOPE BPN will choose one of these PID services as its preferred PID mechanism. The HOPE System will host a dedicated resolver for this preferred service. The CP that wants to make use of this HOPE service must first register an institutional PID at this preferred PID service, before they can start using the HOPE PID service. Next to hosting the PID service, HOPE will also provide tools to help CP add PIDs to their objects and metadata records (for example, a small stand-alone command-line tool that will simplify the addition of PIDs to objects/records in the local system with a JDBC compliant database). When the CP adds a dedicated PID field to their metadata, this tool will automatically populate these fields and register the PID at the HOPE Persistent Identifier Service.

So due to work-flow concerns the consistent usage of PIDs in the local CP system is deemed preferable. This solves many long-term integration problems for aggregation projects like HOPE. These issues are amongst others: duplicates' detection, dealing with orphaned items due to changed identifiers, problems with incremental updates, broken links throughout the system, etc. The PID based work-flow will also remove a lot of architectural bloat from the HOPE system and make the high-level work-flow much more transparent.

5. HOPE Aggregator

This section describes the D-NET Toolkit on which the *Aggregator* will be based and identifies and analyses the functionalities to be used in HOPE. Below a diagrammatic representation of the module can be found.



The D-NET Toolkit

The Aggregator module is realized by means of the D-NET Software Toolkit (for more info: http://www.d-net.research-infrastructures.eu).
The D-NET Toolkit offers a service-oriented framework where developers can build applications by combining a set of D-NET services.
Furthermore, the framework allows for the addition of new service typologies, in order to introduce new functionality, whenever this is required and without compromising the usability of other components. D-NET provides a *data management service kit*, whose services implement functionality for the gathering, manipulation and provision of XML records exported by a set of content providers. Such services have been realized and added in the years to meet the particular requirements arisen when facing new challenges in different application domains (e.g., DRIVER project, EFG project, OpenAIRE project). Most importantly, they are designed according to two engineering principles:

Modularity: services provide minimal functionality and exchange long lists of information objects through the ResultSet mechanism (by relying on a ResultSet Service instance or by implementing natively the interface functionality of the ResultSet Service) so that they can be composed with others to engage in complex data management workflows.

Customizability: services should support polymorphic functionalities, operating over XML records whose data model, i.e., XML format, matches a generic structural template. For example, the D-NET index service is designed to be customizable to index records of any XML format.

As mentioned above, the D-NET framework enables the combination of services into workflows, to obtain complex and personalized data processing operations. To this aim, the services are designed to exchange XML records through mechanisms offered by D-NET ResultSet Services. The service manages *ResultSets*, i.e., "containers" for transferring list of files between a "provider" service and a "consumer" service. Technically, a ResultSet is an ordered lists of files identified by an *End Point Reference* (called EPR, the Web Service EPR standard describes the location of a resource on the Internet), which can be accessed by a consumer through paging mechanisms, while being fed by a provider. D-NET services can be designed to accept or return ResultSet EPRs as input parameters or results to invocations, in order to reduce response delays and limit the objects to be transferred at the consumer side to those effectively needed. For example, while the response to a full-text query may consists of thousands of rank-sorted results, the consumer often requires to access tens of them.

D-NET services for realizing the HOPE Aggregator

The HOPE Aggregator has three main functional objectives:

1. collecting the data from content providers (harvesting, transformation and storage);
2. curating the records (editing, cleansing, enrichment);
3. disseminating the records to third-party systems (pull or push)

In the following sub-sections we describe the D-NET services to be used for the realization of the three objectives and the implementation of the HOPE Aggregator. We shall see that, in order to satisfy HOPE requirements, in some cases D-NET services will have to be extended and new ones will have to be realized.

Collecting the data

Typically, the content providers will export their metadata in the form of XML records through OAI-PMH protocol APIs. The Aggregator has the task of administrating a set of "authorized" content providers in order to harvest their records, if necessary transform them into the HOPE metadata schema, and store them locally. To this aim, the following D-NET services will be combined into a data-set workflow:

Content Provider Manager Service. Content providers are "registered" to the Aggregation system, with a "profile" that describes their location and typology (currently "OAI-PMH compliant" and "remote FTP folders"). The service offers user interfaces for the registration (by content provider administrators) and the subsequent administration of the content providers and their data-sets (by HOPE aggregator administrators, who can fire harvesting, manage transformation mappings, etc).

Harvester Service. An Harvester Service can execute the six OAI-PMH protocol verbs, and communicate with a given data source registered to the system. In particular, the verb ListRecords fetches from the data source the metadata records of a given metadata format (e.g., oai_dc) and returns the EPR of a ResultSet that contains them.

File downloader Service. A File Downloader Service can import all XML files in a given local/remote file system folder. In particular, a "download" call returns an EPR of a ResultSet that contains such files.

Transformator Service. A Transformator Service is capable of transforming metadata records of one data model into records of one output data model. This service is responsible for harmonization of the metadata records. The logic of the transformation, called "mapping", is expressed in terms of a rule language offering operations such as: (i) field removal, addition, concatenation and switch, (ii) regular expressions, (iii) invocation of an algorithm through a Feature Extractor Service, or (iv) upload of full XSLT transformations. User interfaces support administrators at defining, updating and testing a set of mappings. A transformation request is thus composed by: input metadata format, EPR of input ResultSet, output metadata format, reference to the mapping to be applied. If the mapping is not available, the transformation is left pending until HOPE aggregator administrators will provide one. The result of a transformation is the EPR of a ResultSet that contains the generated metadata objects.

MDStore Service. HOPE XML metadata records collected (or obtained by transforming collected records) from content providers are aggregated into MDStore Services. An MDStore (factory) Service manages a set of MDStore units capable of storing metadata objects of a given metadata format. Consumers of the service can create and delete units, and add, remove, update, fetch, get statistics on metadata records from-to a given unit. A given MDStore unit is fed by passing the EPR of the ResultSet containing the incoming records, and when accessed returns an EPR to the ResultSet containing the output records.

Curating Authority Files and metadata

HOPE administrators (often a group of "experts in the field" selected across the content providers) may be willing to perform further semi-automatic cleansing activities. To this aim, the following D-NET Services will be used:

Authority File Service. The Authority File Service implements functionality for "curating" a set of *authority files*, intended as sets of authoritative metadata records. Note that authority files are typically kept, fed, administrated separately from the core data that depend on them, which has to be kept synchronized to the possible updates occurring to the authority files of reference. In particular, administrators can:

- create an authority file, by providing the relative metadata data model,
- create, delete, edit metadata records in it,
- use algorithms for the identification of candidate pairs of duplicate records,

- "merge" a pair of candidate records into one, and "split" metadata records, i.e., obtaining two records from one.

When finished, administrators can "commit" changes and generate a new version of the authority file. Each version is accompanied by a *report file*, which contains the list of merge or split operations committed in the file and that can be exploited by consumers (D-NET services or external applications, see Metadata Editor Service) to upgrade the data in data-sets making use of the authority file according to the latest updates. Consumers can feed an authority file with new metadata records (by sending an EPR of a ResultSet with new metadata records) or access an authority file, which is returned into a ResultSet through an EPR.

Note: *which sets of XML metadata records and which subparts of such records will be identified as authoritative in HOPE will be decided once the HOPE Metadata Schema will be defined.*
Extra: *authority files, as well as report files, may be accessed and exploited also by content providers to improve the quality of their data. Authority files will need to make use of PIDs in order to be accessible in an unequivocal fashion across the HOPE system and beyond.*

Metadata Editor Service. The service offers functionality for the manual or automated editing of the XML metadata records stored into the MDStore Service units. Manually, HOPE administrators can search, add, remove and edit the records in MDStores. Automatically, the report files from Authority Manager Services are processed so that changes can be propagated to the records involved.
Extra: An additional functionality will have to be implemented, to satisfy the following HOPE requirement: the HTTP address of the thumbnails generated by the HOPE Shared Object Repository will have to be added to the HOPE XML metadata records describing the relative objects in a second stage, through a remote request. The functionality will enable authorized applications to invoke an update operation (possibly bulk) over one given field of one given record (by identifier) with one given value (likely the functionality will be generalized to an arbitrary number of fields of the same record).

Disseminating the metadata

HOPE XML metadata records, which are continuously collected and curated as explained above, will be disseminated through different protocols and possibly different XML formats. To this aim, several service workflows will be constructed, one for each typology of data export. Whenever the data requires a transformation into another format, the Transformator Service will be involved: this is the case for the Europeana Metadata Schema (EDM) and for the records to be exported to third-party systems, which generally accept records matching specific import formats. In addition, the following D-NET services will be used:

Index Service. An Index (factory) Service manages a set of Index units capable of indexing metadata records of a given data model, i.e., XML format, and replying full-text CQL queries (*Contextual Query Language*, http://www.loc.gov/standards/sru/specs/cql.html) over such objects. Consumers can

feed units with records, remove records or query the records. The Index Service replies to CQL queries by returning the EPR of a ResultSet that contains the result. Moreover, the service supports advanced full-text highlighted search and faceted browsing functionality. The Service is implemented on Solr (*Solr Apache Lucene Project,* lucene.apache.org/solr/).

Search Web service. A Search web service offers an SRW/CQL (*Search/Retrieval via URL*, http://www.loc.gov/standards/sru) interface accepting a CQL query Q and a metadata format, to run Q over the Index units matching that model. To this aim, queries are routed to the right Index Services; the responses, when more than one Index Service is involved, are then "fused" and pushed into a ResultSet, whose EPR is returned as result. Note that for performance reasons the Search web service memorizes a cache of the Index units available, kept up to date by subscribing to the creation and removal of Index units in the system.

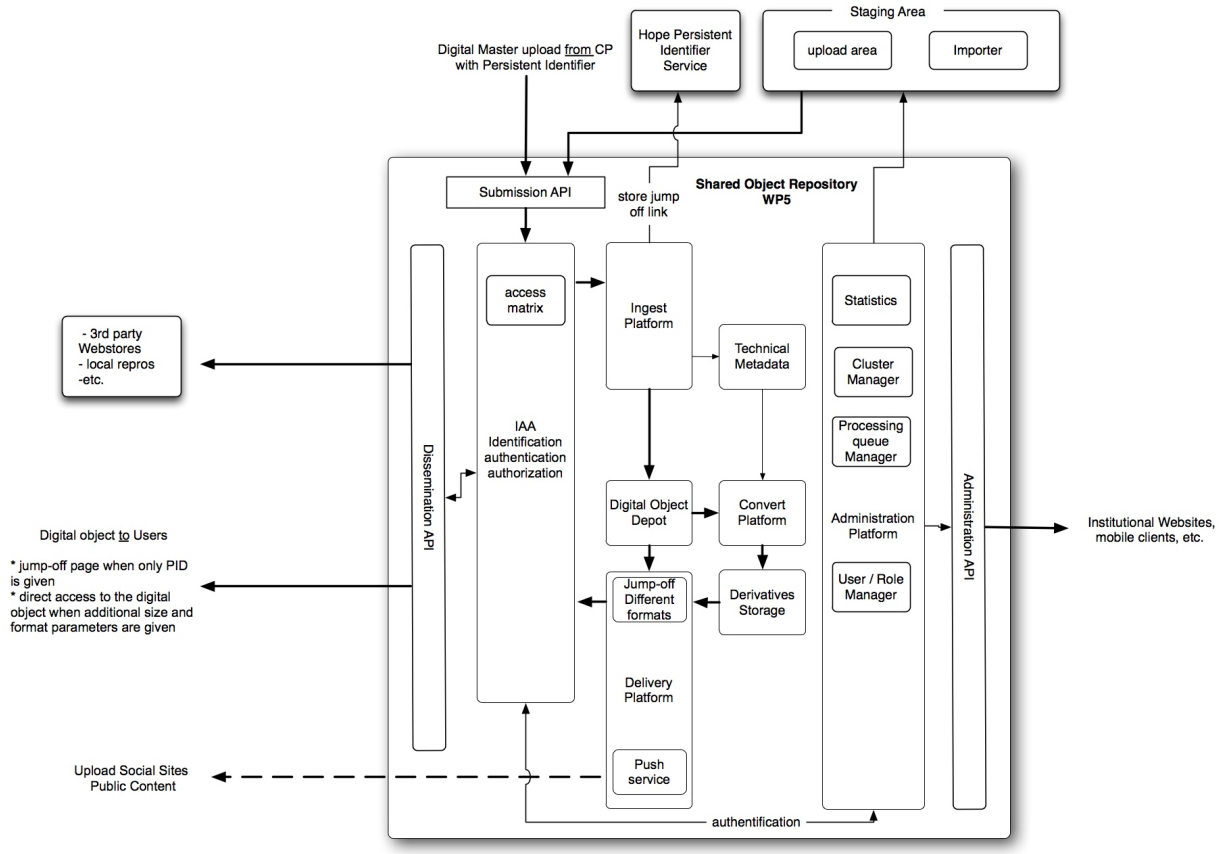
OAI-PMH Publisher Service. An OAI-PMH Publisher Service offers OAI-PMH interfaces to third-party applications (i.e., harvesters) willing to access metadata objects in the MDStore units. To this aim, the service dynamically discovers and exposes through the getFormats verb the list of metadata formats currently available in MDStore units and offers a getRecords operation over all the MDStore units hosting such records.

Export Service. **A new D-NET Service** will have to be realized, capable of exporting XML records from MDStore units to known web sites, such as YouTube, Flickr, Google and Scribd. For certain export services interaction with the SOR is required. Through the SOR dissemination API, the export service is able to determine which derivatives for a digital master is available and can determine the most appropriate format to be exported to the external websites. Preferably, the Export Service will not fetch the derivatives from the website, but construct the appropriate SOR url for external site to retrieve the derivative.

Metadata Synchronization Service. Having a continuous feedback loop between the Aggregator and the content providers is one of the value propositions of the HOPE system. Especially, named entity recognition, the addition of harmonized geographical references, data harmonization, and replacing literals with authoritative URLs are important to CPs. **A new D-NET service** will realize this feedback loop and make it possible for HOPE and the CPs to become an integral part of the Linked Open Data cloud.

6. HOPE Shared Object Repository

This section describes the design considerations and proposed components for the Shared Object Repository. The SOR plays a critical role in the d2d process to make access to the digital masters and their derivatives more transparent to the user. In the future, the SOR can also play a critical role in the digital preservation of the digital masters.
Below a diagrammatic representation of the Shared Object Repository can be found.



Design considerations

- Descriptive metadata (of the digital objects) are for search and discovery. They are not functional or necessary for the operation of the repository. These metadata are produced and managed outside the digital object repository. The repository is descriptive metadata-agnostic.
- In order to promote the open web character of the SOR, all interaction with repository should go through Web-based APIs. Currently, three access types are identified: Submission, Dissemination and access to the Administration information.
- Converting digital master files to derivatives (i.e. different formats and sizes) is a function of the repository. The Aggregator is also capable of doing this function, but the Aggregator needs to fetch the digital master file from the repository first. It is better if the Aggregator gets the preview from the repository. Based on the availability of the PID in the descriptive metadata, the Aggregator should be able to construct the dissemination API URL for thumbnail retrieval to the metadata in Aggregator storage.
- Streaming **previews** of films will not be supported by the repository. The previews will be put on YouTube and the link provided to the content provider and/or aggregator. The jump-off page if applicable will also contain links to the previews stored outside the SOR.
- Streaming of **digital objects** (master or derivative) will not be supported by the repository. Once a user decides he/she wants to see the whole film (after having

seen the preview), a download copy will be made available to the user via the delivery platform.

- The repository will not natively contain a payment module. In order to recover cost for maintenance and organizational operation of the HOPE Repository interaction with third party web stores and local reproduction departments is envisioned. These parties would interact with the Dissemination API to acquire a digital object from the delivery platform after the payment has been completed. A scenario where a temporary link is created by the SOR to download the object via the Delivery Platform is also considered. The access to this temporary link will be managed through the dissemination API.
- The SOR is expected to grow rapidly in the coming years both in number of objects and number of users. Therefore scalability and the ability to replicate is of paramount importance. The content of the three storage components - Digital Object Depot, Technical Metadata Storage, and Derivative storage - should be easily replicable to different nodes in the cluster. In addition, the use of a Content Delivery Network provider should be investigated for the license-free objects to reduce the load and bandwidth requirements on the main SOR. Alternatively -- because of the single access design principle through the dissemination API -- the use of geographically distributed caching proxies is also considered to reduce the load on the main SOR. The latter option would transparently interact with the IAA module.
- The SOR should be able to run as the master in a cluster - like it would in the HOPE System - but it should also be able to be installed on the local infrastructure of the content provider as a stand-alone system. Ideally, a remotely installed SOR should be able to be added to the cluster of the central HOPE Repository streamlining the synchronization even further. This requirement is primarily geared towards enabling the CP to have their own digital repositories. This is a wish that is shared by many CPs in the HOPE network.

Shared Object Repository Components (SOR)

The following sections contain a short description of the functionality of each component in the SOR diagram.

APIs

The three public APIs are the only access mechanism to the SOR. Through the use of simple XML based APIs we hope to make the use and integration of SOR in existing workflows as easy as possible. Three access types have been identified: submission, dissemination and administration. Since the APIs are web-based it means they will take full advantage of asynchronous and concurrent processing of requests.

Submission API

The submission API is responsible for receiving a submission request for storing a digital master in the SOR. The XML processing instruction also contains an option to

send a delete request for the digital master. See also Appendix: Sample XML Processing Instruction The access information will be a controlled set of license options. In the Administration platform, the CP also has the possibility to add group based permission to the whole collection.

- request: persistent identifier, mime-type, access information, location of the digital object (this can be on local disk, URL, or as part of the HTTP post), API-access key, and a content checksum of the original object. The checksum can be used as a mechanism to ensure the correct item is transferred to the SOR. In the Ingest Platform section, this option is further elaborated upon.
- response: XML with status-code

Dissemination API

The dissemination API is the single point of access for all requests for digital objects in the SOR for both human web-users and machine-to-machine interaction. When a request is made to this API with only a persistent identifier the response will be a jump-off page (either as HTML, XML, etc) that contains links to the different available derivatives for the digital object and the license conditions. To some links the description and license information alone will be available, depending on the access restrictions. The links consist of the API base-URL plus the PID, format, size and if applicable the API-access key. The PID refers to the master object that is submitted via the submission API. The derivatives are all linked to the master PID. The sizes and formats of the derivatives are stored as part of the Technical Metadata of the master object identified by the PID.

- request: PID, size, format, API-access key, (output format: JSON, XML, HTML, etc.)
- response: jump-off page when only the PID is given. When the other parameters are given direct access to the object is supplied.

Administration API

The administration API will consist of different components that give access to the different parts of the Administration platform. The rendering layer of the Administration platform will use the same API. A subsection of this API will be made available to partners so they can use this information on their local websites. For authentication a web-services/API key will be made available via the user/role management component.

IAA: Identification, Authentication, Authorization

The SOR has an identification, authentication and authorization system. This is necessary to act on access restriction rules, which apply to categories of users in combination with types of usage of digital objects. This feature makes the repository a "trusted repository": the archival collections entrusted to the CPs are not always publicly accessible due to the privacy of personal papers. The repository should enforce restrictions on access in a very secure way. Digital objects that are publicly available will be made directly accessible via the dissemination API when besides the

persistent identifier also the format and size parameters are given. Without the format and size parameters, the jump-off page for the requested object is returned.

The IAA system will support both web-services key (wskey) and user/password based authentication. Based on the access matrix and the access information from the Technical Metadata, the IAA system will determine if and to which formats the requester has access to. The IAA system will authenticate all access to the SOR and will be role-based. The roles will be specified in the Access and Use Specifications and Use-Case documents. The role of the user is also part of the access matrix.

Ingest Platform

The Ingest Platform will validate the submission request from the submission API. The validation might possibly also include virus checking of the digital object. Although for security reasons this check could also be done earlier in the workflow. After validation the ingestion platform adds the request on the processing queues for storage of the object and the technical metadata. The technical metadata will also contain a checksum of the digital master. The digital master is stored with the checksum as the identifier in the Digital Object Repository. This will ensure that no duplicates will be stored in the SOR and that updating the digital master attached to the persistent identifier is a straight forward replacement. It also means that multiple persistent identifiers can refer to the same object stored in the SOR. In addition, the checksum is used to make sure that the item has arrived uncorrupted via the web. It can also be used as a quality check to ensure the object is correctly stored and preserved by the SOR.

Administration platform

The access to the administration API should be handled by the IAA component. The widgets should also be accessible from the institutional websites. Since not all institutional websites will have been built in JAVA, the XML API will have as an additional benefit that it will be easy for partners to create views on their data in the SOR in their preferred technologies.

Some of the envisioned functionality of the Administration platform (note that this list is not exhaustive):

- Overall statistics:
  - Number of data-sets
  - Number of providers
  - Number of digital masters
  - Number of derivatives
  - Access request per data-set, provider, per object
- Monitoring:
  - progress of import process,
  - progress of conversion process,
- Data-set settings:

    - determine access privileges,
    - determine which derivatives are created,
    - determine which derivatives are pushed to social websites
- Role/Group management
    - add users to groups
    - determine group based access privileges

Technical Metadata storage

For the SOR to manage a digital object correctly some basic technical metadata must be supplied during the submission phase. For example, a persistent identifier, mime-type of the object, and license/access information. This information is used by various other components of the SOR to manage the workflow. In addition, a CP provided checksum during submission is also being considered. This checksum will be used for duplicates detection, quality assurance (whilst receiving the object and during storage), and as storage id in the digital object depot.

The technical metadata will be stored in a replicable document database. This database is an integral part of the SOR. Because the Technical Metadata storage must be able to function in a cluster the information must be redundantly available. The ingestion platform is responsible for storing the initial record. Several components can update a technical metadata record: administration platform, processing queue (i.e. digital object is stored, derivative stored, object flagged for push to social websites, etc.).

Digital Object Depot

The digital object depot is where all the digital masters are stored. The store will have to be replicated to provide redundant storage. The stored digital object is identified by the content checksum. The checksum is stored as part of the technical metadata record for each digital master.

Convert Platform

The Convert Platform should be able to handle a wide variety of formats and create derivatives in most current web-standards. Since the creation of the derivatives - especially from movies and HD master files - is computationally very expensive parallelization of these processes is an absolute requirement. The convert platform should therefore seamlessly interact with Processing Queue Manager to acquire transformation tasks and be able to run stand-alone on different nodes in the cluster. In order to make the Convert Platform easily extendible, a plug-in-based approach for the different converters is the preferred option.

Derivative storage

The derivative storage is responsible for managing the derivatives of the digital master objects that are stored in the Digital Object Depot and are created by the Convert Platform. Although the derivative storage is a separate component in the High Level Design, it is most likely that both the digital master and the derivatives are

stored in the same storage mechanism. For the moment, we assume that the SOR will create derivatives for both Video and Image digital masters.

The Derivative Storage should interact with the Cluster Manager. It is likely that the Derivative storage will consist of multiple shards that need to have a single interface to query for and insert derivatives. This multi-node setup of the storage will ensure high throughput for Delivery platform and Convert platform. In a distributed environment the storage and retrieval of derivatives could easily become a bottleneck.

Cluster Manager

The cluster synchronization/replication manager is responsible for distributing the digital object and technical metadata across the cluster. Although this functionality will most likely be supplied by the technical storage solutions that will be chosen for the SOR, it is still an important part of the SOR. It is a requirement that these storage solutions have some kind of API that makes it possible to integrate information on the state of the cluster in the Administration Platform API.

Processing Queue Manager

The Processing Queue Manager should enable a transparent work-flow between the different components of the SOR. The benefits of an Event Driven Architecture where the components interact with each other through queues is that it becomes much easier to distribute work in the cluster (e.g. use cloud-based solutions to dynamically scale up processing capacity during peak-times) and to use state-based work-flows to prioritize tasks on the queue.

Delivery Platform

An important function of the repository is the interfacing platform responsible for delivering digital objects from the repository upon request (directly to end-users or to external systems). The delivery platform should be capable of accessing derivatives of the master digital copy into a wide variety of formats from the derivative storage. The jump-off page is generated from the Technical Metadata record of the requested PID. It will also need to interact with the IAA to determine if the requested object is available based on the requester's access privileges.

The Delivery platform will be a web application server that will most likely also need to be clustered. Due to the stateless nature of the Dissemination API scalability can easily be achieved through horizontal scaling of the web application servers.

The delivery platform also contains a push service that is able to push derivatives to social sites like Flickr and Youtube. The CP has the possibility to specify in the Administration platform which objects will be pushed to the external services.

Related Components

The related components are part of the SOR ecosphere, but are not considered core-functionality. These components interact with the SOR and provide valuable additions to the work-flow and manageability of the HOPE system.

- **HOPE Persistent Identifier Service:** The Ingest Platform will interact with a number of Persistent Identifier Services, like for example Handle. The main purpose of this interaction will be to make sure that the right information is added to the properties of the PID, so that the ingestion-process requires a minimal amount of additional Technical Metadata. As stated before, only objects with a persistent identifier will be able to be submitted to the SOR. HOPE will also offer its own dedicated Persistent Identifier service to make it easier for CPs to implement PIDs in their system without having to maintain a PID service themselves. The CPs are responsible for requesting an institutional PID, before the HOPE Persistent Identifier Service can host their PIDs for them.
- **3rd party Web stores and reproduction units:** The web stores and reproduction units will interact with the delivery platform via the dissemination API to transparently handle transaction for users that want to buy an object that resides in the SOR. They can make use of web-payment services like Paypal.
- **SOR submission tool:** The SOR submission tool is a small command-line tool that will be supplied to the CP. It will process the XML processing instructions and make the API calls to the SOR ingestion API. The SOR submission tool thus takes care of all the boiler-plate of using the Ingestion API and provides convenient local validation of the correctness of the XML processing instruction file.
- **Staging area (previously known as pre-ingest area):** Since not all content providers are able to store all digital objects online or send them via http, a scenario with FTP upload should be supported as well. The basic idea is that the CP uploads the objects to the staging area together with an XML processing instruction. This instruction contains all the parameters to construct calls to the Submission API. From the Administration platform, the CP should be able to trigger a run of the importer that reads the processing instruction and turns them into Submission API calls. The CP should be able to track the progress of the import via one of the Administration Platform widgets.
  The benefit of creating a staging area early on in the project, is that the CPs can start processing their collections over a long period of time as opposed to creating bottlenecks before go-live time.

CONCLUSION

The focus in the High Level Design, Architecture and work flow of the HOPE System is on providing a transparent discovery-to-delivery process for the users and a simple, maintainable work-flow[4] for the HOPE content providers. The system is designed with future maintainability, sustainability and scalability in mind. By the completion of the project the system should be operational and should be able to sustain itself afterwards through the various cost-recovery strategies developed in the WP7 HOPE Exploitation Plan.

[4] The work-flow for HOPE Content Providers is fully and practically explained in the "HOPE Manual for Content Providers" [http://igwiki.peoplesheritage.eu].

Appendix: Sample XML Processing Instruction

Below you find the first draft for the processing instruction that the HOPE SOR submission tool needs to submit each object to the SOR. The SOR submission tool will be a small standalone java tool that will be supplied by the HOPE to the Content Providers to simplify the submission to the SOR.

```
<hope_sor api_key=”123/456”>
        <object action=”add”> <!-- default: add but delete other option -->
                <pid>..../1066/1</pid> <!-- persistent identifier -->
                <mime-type>image/jpeg</mime-type> <!-- standard http mime-types -
->
                <access>free</access>
                <location>/Volumes/hope/coll1/image.jpg</location>
                <checksum>42eac3764dea6aabf4c92add6beb636a</checksum>
        </object>
        ..... <!-- more objects -->
</hope_sor>
```

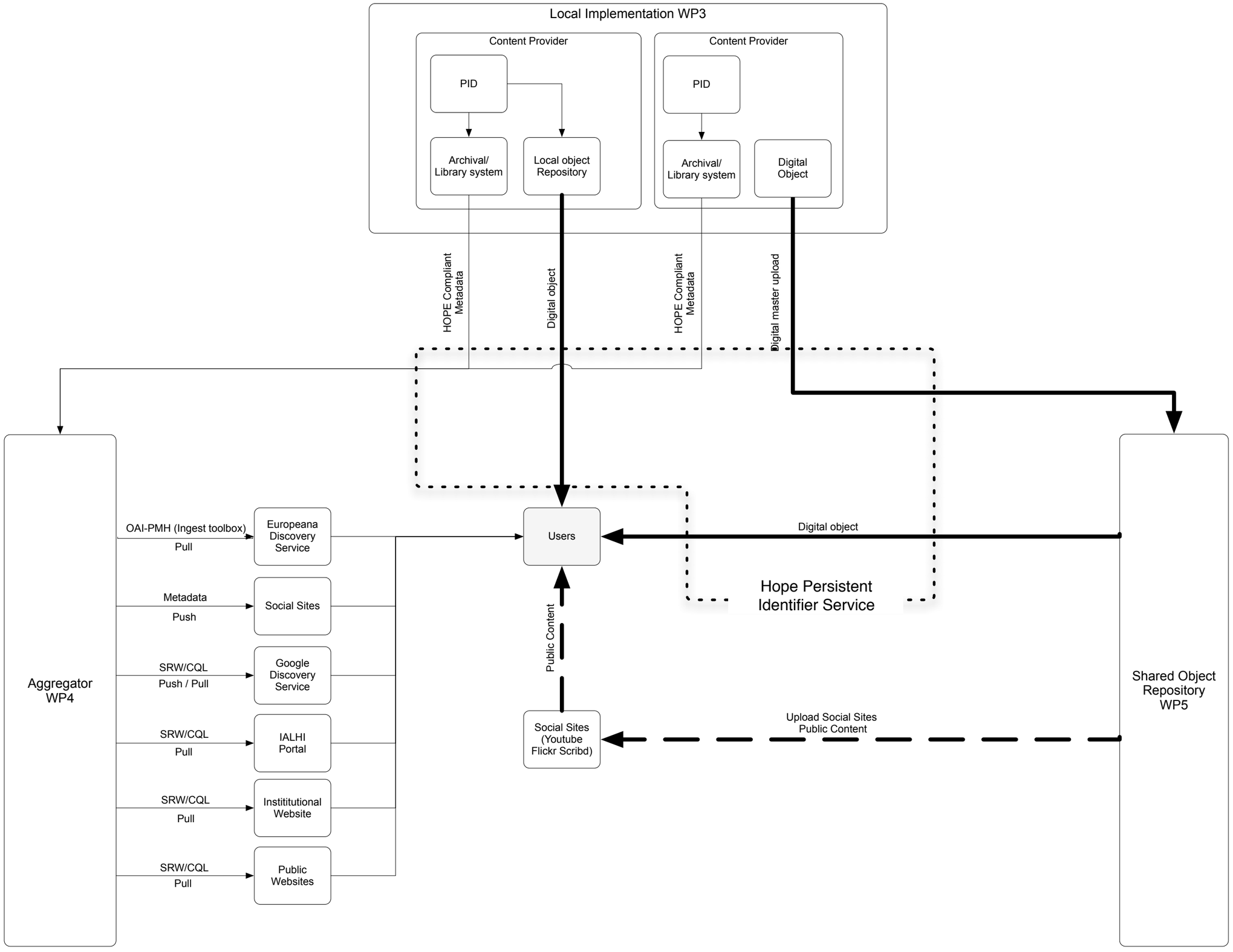
Appendix: High Level Design Diagrams

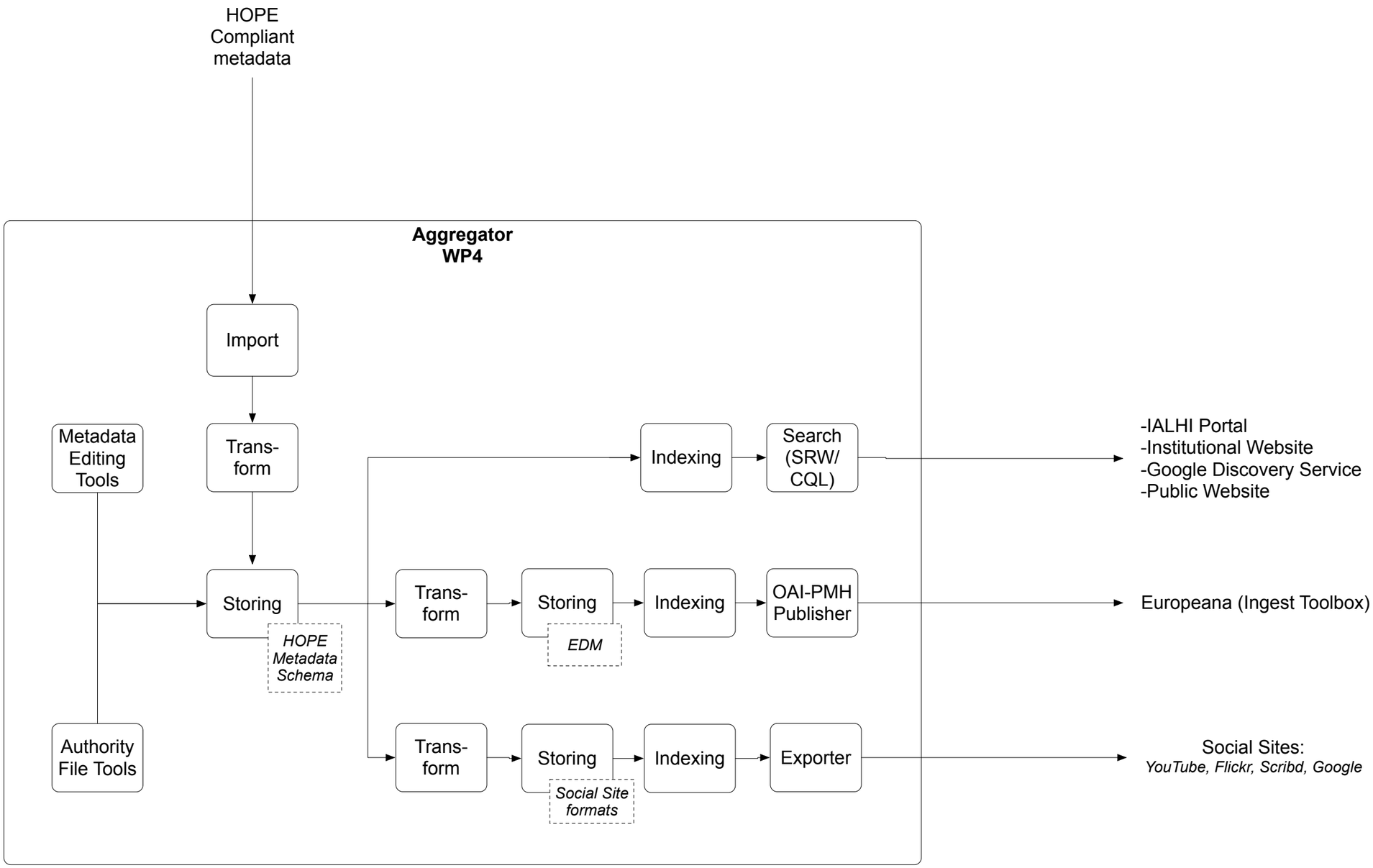
See next set of pages

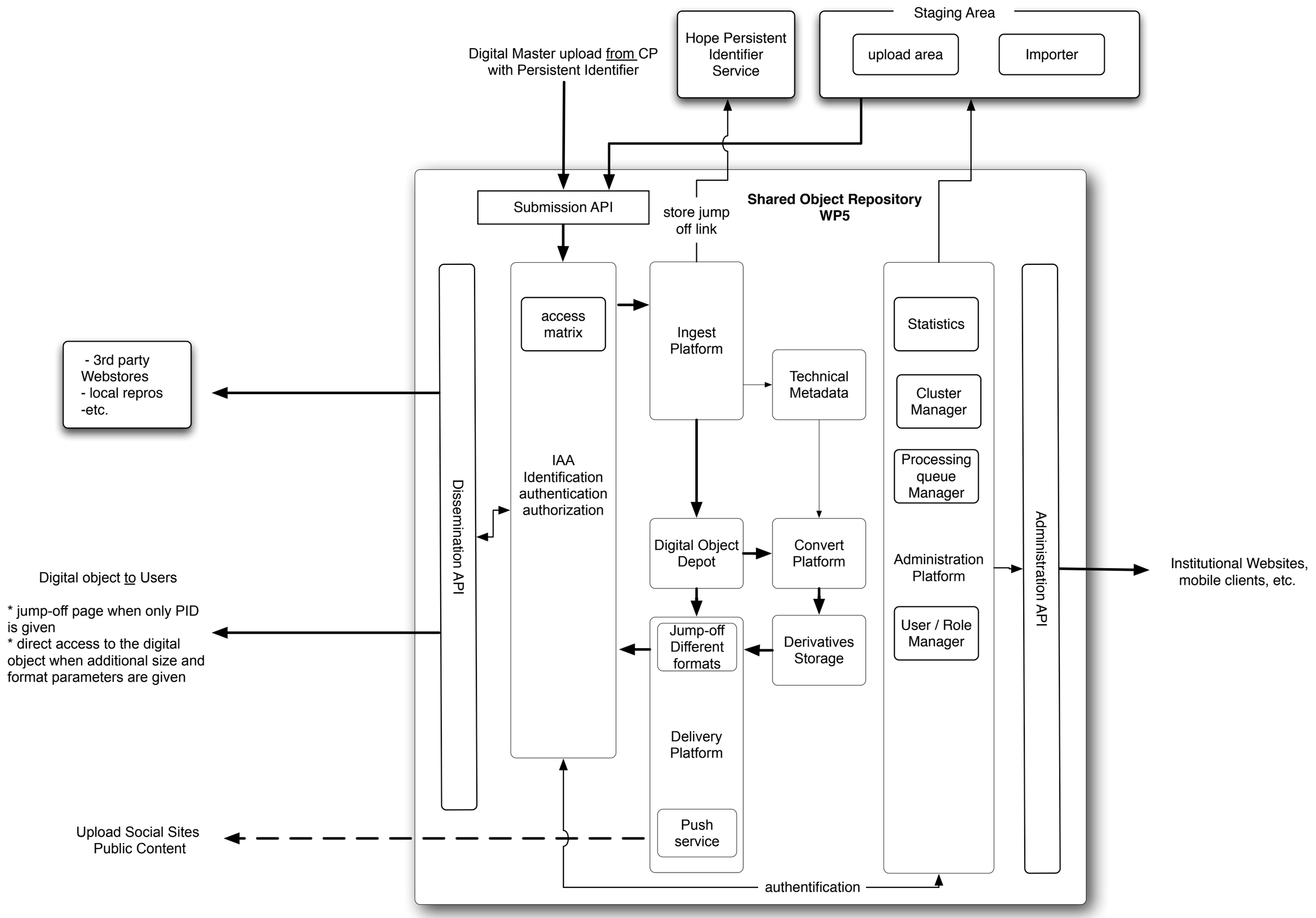
Appendix: HOPE Glossary

See next set of pages

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


Local Implementation WP3
Content Provider
PID
Archival/
Library system
Local object
Repository
Content Provider
PID
Archival/
Library system
Digital
Object
HOPE Compliant
Metadata
Digital object
HOPE Compliant
Metadata
Digital master upload
Aggregator
WP4
OAI-PMH (Ingest toolbox)
Pull
Europeana
Discovery
Service
Metadata
Push
Social Sites
SRW/CQL
Push / Pull
Google
Discovery
Service
SRW/CQL
Pull
IALHI
Portal
SRW/CQL
Pull
Instititutional
Website
SRW/CQL
Pull
Public
Websites
Users
Digital object
Hope Persistent
Identifier Service
Public Content
Social Sites
(Youtube
Flickr Scribd)
Upload Social Sites
Public Content
Shared Object
Repository
WP5


HOPE
Compliant
metadata
Aggregator
WP4
Import
Metadata
Editing
Tools
Trans-
form
Indexing
Search
(SRW/
CQL)
-IALHI Portal
-Institutional Website
-Google Discovery Service
-Public Website
Storing
HOPE
Metadata
Schema
Trans-
form
Storing
EDM
Indexing
OAI-PMH
Publisher
Europeana (Ingest Toolbox)
Authority
File Tools
Trans-
form
Storing
Social Site
formats
Indexing
Exporter
Social Sites:
YouTube, Flickr, Scribd, Google


Digital Master upload from CP
with Persistent Identifier
Hope Persistent
Identifier
Service
Staging Area
upload area
Importer
Submission API
store jump
off link
Shared Object Repository
WP5
- 3rd party
Webstores
- local repros
-etc.
access
matrix
Ingest
Platform
Technical
Metadata
Statistics
Cluster
Manager
Processing
queue
Manager
Administration
Platform
User / Role
Manager
Dissemination API
IAA
Identification
authentication
authorization
Digital Object
Depot
Convert
Platform
Derivatives
Storage
Jump-off
Different
formats
Delivery
Platform
Push
service
Administration API
Institutional Websites,
mobile clients, etc.
Digital object to Users
* jump-off page when only PID
is given
* direct access to the digital
object when additional size and
format parameters are given
Upload Social Sites
Public Content
authentification

HOPE GLOSSARY - V2.1

This is the HOPE glossary, providing definitions of acronyms and abbreviations and a terminology. There are many terms used in HOPE which need to have well-defined meanings, and these are defined in this document.
HOPE takes its terminology from the cultural heritage sector, the research sector and the computer sciences. Same terms have sometimes different meanings in each sector and are therefore ambiguous across different disciplines (e.g., traditional archives, digital libraries, open access repositories, research data centres).
The approach taken is to opt for terms in use by the cultural heritage sector wherever possible in order to allow for specificity and to avoid terms that are overloaded with meaning in several disciplines, so as to reduce ambiguity.

The HOPE Glossary is a work in progress with Google Docs based ongoing work. This Version 2.0 has been cut 21 March 2011. With HOPE implementation oriented Technical Documentation in the making, some glossary terms have been moved to or will be published in specific glossaries that will appear as appendices to such Technical Documentation.

In this glossary we used different sources. When a definition was taken from or based on an external source, the (clickable) source is mentioned.

- Glossary of the Society of American Archivist (SAA)
- OAIS Reference Model (Blue Book, January 2002)
- HOPE Glossary Version 0.2 (HOPE Consortium, July 2010)
- PREMIS 2.1

Definitions newly developed are indicated with “HOPE” as source.

For cross-referencing “See” and “See also” are in use. The “Use” and “Use for” signify preferred terms. In both cases the glossary terms referenced to or preferred are in italics with first letters capitalised.

We grouped the glossary terms as follows:

1. HOPE System
2. Metadata
3. PIDs and Identifiers
4. Digital Objects
5. IPR
6. Dissemination

Abbreviations

d2d: discovery to delivery

DoW: Description of Work

EDM: Europeana Data Model

ESE: Europeana Semantic Elements

HOPE: Heritage of the People's Europe

IALHI: International Association of Labour History Institutions

IPR: Intellectual Property Rights

OAI: Open Archives Initiative

OAIS: Open Archival Information System

Terminology

1. HOPE System

Term	Meaning	Comment
Collection	A set of items with one or more common factors, such as material type, author, publisher, provenance, and/or subject. In HOPE, Collections are provided by CPs in the form of metadata records and, if available, *Digital Objects*. Collections are used as the basis for submission and management of records and objects; and serve as a key access point for end users. As such, collections must meet certain technical requirements. Note that metadata records belonging to one Collection, may be submitted to the *Aggregator* as one or more *Data Sets*.	Source: HOPE Use for: *Sub-Collection* See also: *Data Set* *HOPE Themes* Note that the HOPE definition of Collection differs from the definition commonly used in the professional community, which holds a collection to be the entire holdings of a single repository.
Discovery Service	A web portal, which enables the discovery, identification, and selection of materials through searching and browsing functions.	Source: Glossary v.1
Discovery to Delivery (d 2 d)	A process that offers all appropriate options to the unassisted information seeker on the web. The journey between discovery and delivery is accomplished with a variety of differing technologies and processes, many of which fall under the responsibility of different	Source: Glossary v.1

	providers (discovery services, aggregators, repositories, etc.).	
HOPE Aggregator	The system that harvests, stores, and disseminates *Descriptive Metadata* supplied by CPs. The Aggregator enables harmonisation and enrichment of the metadata and provides a *Search API* for use by the *IALHI Portal* and CP institutional websites.	Source: HOPE
HOPE Best Practices	Best Practices are generally-accepted, informally-standardized techniques, methods, or processes that have proven themselves over time to accomplish given tasks and unlike standards are often highly context dependent. In HOPE, Best Practices are used to standardize practice across the BPN with the aim to increase interoperability and to enhance the quality of service. In the early phases of the project, Best Practices are gathered from the professional and technical fields to feed into system development. In later phases, HOPE will publish Best Practices based on its own experience.	Source: Wikipedia (with editing)
HOPE Content Provider (CP)	A HOPE partner with social history collections which provides metadata and *Digital Objects* to the *HOPE System*.	Source: Glossary v.1
HOPE PID Service		Source: HOPE
HOPE Search API	The Application Programming Interface that defines the available methods of the *Search Web Service*. Software systems performing searches on the HOPE Collections can call the Search Web Service using the API via REST or SOAP protocols.	Source: Hope See also: *Search Web Service*
HOPE Shared Object Repository (SOR)	The shared *HOPE-Compliant Digital Object Repository* used by some CPs for the ingest, storage, management and delivery of their *Digital Objects*.	Source: HOPE See also: *HOPE-Compliant Digital Object Repository*
HOPE Social History Resource	The selection of *Collections* brought together by the HOPE CPs with the aim of making a coherent and rich social history resource available through *Discovery Services*.	Source: Glossary v.1 See also: *Collection*
HOPE Support	A group of specialists selected from among the members of the consortium, in charge of	

Team (HOPE-ST)	the assistance of the CPs to support the implementation of the supply chain.	
HOPE System	The set of interdependent entities (CP local information systems, *HOPE Aggregator*, *PID Services*, *HOPE Shared Object Repository*, *Discovery Services*) forming the integrated whole—called the HOPE System—with the purpose of executing the functions defined in the HOPE high-level design.	Source: Glossary v.1
HOPE-Compliant Digital Object Repository	A digital object repository, digital assets management system, or other network accessible system that is used for the ingest, storage, management, and delivery of *Digital Objects* and that is compliant to a set of agreed minimum functionalities and services within the HOPE system	Source: Glossary v.1 Note that this is distinct from local bibliographic utilities or collection management systems, which support the production, storage, and delivery of Descriptive Metadata.
IALHI Portal	The *Discovery Service*, also known as Labourhistory.net, provided and maintained by the International Association for Labour History Institutions (IALHI) for the social science and history research community. The IALHI Portal will be upgraded during the HOPE project, and IALHI will be asked to provide an official name for the portal at that point.	Source: Glossary v.1
Local Object Repository (LOR)	A *HOPE-Compliant Digital Object Repository* that is maintained by a HOPE CP.	Source: HOPE See also: *HOPE-Compliant Digital Object Repository*
Local PID Service	A PID Service such as ARK, Handle System, etc., locally hosted and maintained by the CP. Permits the creation of PIDs and the binding between the resolve URL and the PID.	Source: HOPE
Search Web Service	A web service provided by the *HOPE Aggregator* enabling the search of metadata records. The service can be accessed by applications via the *HOPE Search API*.	Source: Glossary v.1 See also: *HOPE Search API*
Social Sites	Websites that attract users to share and exchange information, usually for one specific purpose (networking, bookmarking, etc.) or one specific medium (videos, photographs, etc.). Examples include YouTube, Flickr, Scribd, and Facebook.	Source: Glossary v.1

Sub-Collection	Use: *Collection*	
Third-Party PID Service	A PID service, offered by a third party (regional/national/commercial/etc. service) and used by the CP for *PID* creation and *Binding* to the *Resolve URL*.	Source: HOPE

2. Metadata

Term	Meaning	Comment
Administrative Metadata	Data necessary to manage, process, use and preserve digital objects and metadata. Administrative metadata generally includes: *Technical Metadata*, rights management metadata, and preservation metadata. Administrative metadata is stored and managed throughout the entities of the *HOPE System*.	Source: HOPE See also: *Technical Metadata*
Aggregator Processing Instruction	The XML format for exchange of *Descriptive* and *Structural Metadata* about *Digital Objects* between the CP and the *Aggregator*. To be used by CPs that are not able to integrate metadata about digital objects in their collection management system.	Source: HOPE See also: *Structural Metadata*
Archival Finding Aid	*Descriptive Metadata* on the records composing an archival collection. The Archival Finding Aid is generally hierarchic, describing the collection from general to specific, starting with the whole then proceeding to the components (fonds, series, files, and items). Such metadata are usually created and captured in an archival management system. The HOPE data model supports hierarchic description, such as that characteristic of an Archival Finding Aid.	Source: HOPE See also: *Descriptive Metadata,* *File,* *Fonds,* *Item,* *Series*
Authority File	Use: *Authority List*	
Authority List	A controlled vocabulary composed of a set of *Authority Records* on descriptive terms, names, phrases, or similar entries, which enable cataloguers to disambiguate descriptions with similar or identical headings and to collocate objects that logically belong together but that are presented in a different way.	Source: HOPE Use for: *Authority File* See also: *Authority Record*

		HOPE Theme
Authority Record	An entry in an *Authority List* that contains an identifier and the preferred name of the term. Authority Records may also include additional metadata about the term, such as variant names, translations, descriptions, dates, etc.	Source: HOPE See also: *Authority List*
Bibliographic Description	*Descriptive Metadata* on library collection items. Bibliographic Description is formal description providing access to each item and its content. Such metadata are usually produced with a library information system or bibliographic utility. The HOPE data model supports analytic description such as that characteristic of Bibliographic Description.	Source: HOPE See also: *Descriptive Metadata*
Crosswalk	Use: *Mapping*	
Data Set	A group of metadata records accessible to the *Aggregator* from the same entry point. An entry point is a location, either local or remote, identified by a protocol and URI where the harvester can find the Data Set. In HOPE, a Data Set contains homogeneous XML metadata records — metadata records belonging to the same *Collection*, with the same metadata format to be mapped into the same *HOPE Domain Profile* using the same mapping worksheet(s) and serialized in one or more XML files.	Source: HOPE Use for: *Record Group* *Record Set* See also: *Collection*
Descriptive Metadata	Describes the intellectual content of collection materials. Descriptive Metadata are used to facilitate the discovery, identification, and selection of materials. In HOPE, Descriptive Metadata are harvested from local information systems by the *Aggregator* and passed to *Discovery Services*. Descriptive Metadata is gathered for materials with and without related *Digital Objects*.	Source: HOPE See also: *Archival Finding Aid* *Bibliographic Description*
Descriptive Unit	The entity in the HOPE data model that represents a metadata record about one or more materials that are contained by a *Collection*. The Descriptive Unit provides the semantic context for all related *Digital Resources* or for all related child Descriptive Units.	Source: HOPE See also: *Digital Resource*
Digital Resource	The entity in the HOPE data model that accommodates the *Administrative* and *Structural Metadat*a about each *Digital File* composing the *Digital Object* described by a *Descriptive Unit*.	Source: HOPE See also: *Descriptive Unit*

File	An organised unit of documents grouped together either for current use by the creator or in the process of archival arrangement, because they relate to the same subject, activity, or transaction. A file is usually the basic unit within a record series	Source: Isad-G See also: *Archival Finding Aid*
Fonds	The whole of the records, regardless of form or medium, organically created and/or accumulated and used by a particular person, family, or corporate body in the course of thatcreator's activities and functions	Source: Isad-G See also: *Archival Finding Aid*
Granularity	Used to describe the *Levels of Description* making up a hierarchy or the specificity of *Descriptive Unit*.	Source: HOPE See also: *Level of Description*
HOPE Domain Profile	A subset of *Metadata Elements*, borrowed from an domain-specific metadata standard. HOPE provides 5 subsets, one for each domain type and each related with a domain specific metadata standard. The archive profile is based on the APEnet/EAD standard; the library profile is based on the MARC21 bibliographic standard; the audio-visual profile is based on the EN15907 standard; the visual domain profile is based on the LIDO standard; the 'generic' Dublin Core profile is based on the Dublin Core standard. HOPE Domain Profiles are used as an intermediate *Metadata Structure* when mapping local metadata elements to the common HOPE *Metadata Structure*.	Source: HOPE
HOPE Metadata Schema	The XML schema used by the *Aggregator* for validating and storing the metadata harvested from CPs.	Source: HOPE See also: *Metadata Schema*
HOPE Theme	A thematic heading specific to the fields of social and labour history. HOPE Themes are assigned by CPs to HOPE *Collections* or to groups of records within a single collection. HOPE Themes are primarily used to provide uniform cross-language, cross-domain access to the *HOPE Social History Resource*.	Source: HOPE See: *Collection*
Item	The smallest intellectually indivisible archival unit, e.g., a letter, memorandum, report, photograph, sound recording	Source: Isad-G See also: *Archival Finding Aid*
Level of Description	Level of *Granularity* of a *Descriptive Unit* that is part of a hierarchical description. The	Source: HOPE

	designation of the level is generally specific to the collection domain. (E.g. for archival collections, this might include fonds, series, files, and items, while for library collections, series, titles, and issues.) HOPE does not limit the number and type of Levels of Description and can also support idiosyncratic descriptive levels.	Use for: *Level of Granularity* See also: *Descriptive Unit* *Granularity*
Level of Granularity	Use: *Level of Description*	
Mapping	Process of creating links between *Metadata Elements* of two distinct *Metadata Structures*, e.g. between a local, homebrew structure and a *Metadata Standard* or between two metadata standards. A mapping, also called mapping rules, is a specification of such associations between metadata structures. In HOPE, local metadata structures are mapped through one of the five *HOPE Domain Profiles* to the common HOPE metadata structure. The HOPE metadata structure has been mapped to several target *Metadata Structures* including EDM and DC.	Source: HOPE Use for: *Crosswalk*
Metadata Element	A single unit of a metadata set, containing a particular category of information (e.g. 'Date' or 'Creator'). Metadata Elements generally have a name (or label) and a cardinality, specifying whether the element is mandatory and or repeatable. The values of metadata elements may have a controlled semantics and/or syntax.	Source: HOPE Use for: *Metadata Field*
Metadata Field	Use: *Metadata Element*	
Metadata Schema	A formalized description of a *Metadata Structure* using a language such as SGML or XML. For example, MARC/XML is a schema for the MARC21 bibliographic metadata structure.	Source: HOPE See also: *Metadata Structure* *HOPE Metadata Schema*
Metadata Standard	A published document, specifying names, definitions, syntax and/or values of an agreed set of *Metadata Elements*. Metadata Standards may be cataloguing standards (e.g. ISBD), encoding standards (e.g. EAD), and exchange protocols (METS, OAI-ORE).	Source: HOPE
Metadata Structure	A structured set of *Metadata Elements*. A Metadata Structure generally has two components: a defined set of metadata elements, which may include the semantics and syntax of the element, and the structural relationship between these elements. A	Source: HOPE See also: *Metadata Schema*

	metadata structure is a logical structure, which is formalised in a *Metadata Schema*.	
Record Group	Use: *Data Set*	
Record Set	Use: *Data Set*	
Series	Documents arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation or filing process, or the same activity; have a particular form; or because of some other relationship arising out of their creation, receipt or use. A series is also known as a records series	Source: Isad-G See also: *Archival Finding Aid*
Structural Metadata	Describes the internal structure of *Digital Objects* and the relationships between their parts. It is used to enable navigation and presentation of digital objects.	Source: PREMIS (with editing)
Technical Metadata	Describes the physical (as opposed to intellectual) attributes or properties of a *Digital File*. Some Technical Metadata properties are format specific, while others are format independent. In HOPE, Technical Metadata is generally stored and managed in the SOR, LORs, or local information systems.	Source: PREMIS See also: *Administrative Metadata*

3. PIDs and Identifiers

Term	Meaning	Comment
Binding	The association of identifiers and data elements and their storage in a PID service. A binding may include, for example, the association between a PID and a Resolve URL or the association between a PID and a Local Identifier. Several data elements may be bound with a single identifier.	Source: HOPE
Local Identifier	A string that acts as an unambiguous reference to the resource in the context of the local information system.	Source: HOPE See also: *Reverse Look Up*
Persistent Identifier	Use: *PID*	

PID	A character string that is globally unique and permanently identifies a resource within a given context. In HOPE, PIDs are always associated with a resolve URL and should be persistently resolvable on the Internet.	Source: HOPE Use for: *Persistent Identifier*
Resolve URL	A URL associated with a PID in a PID service. In other words, the URL you are redirected to when you perform a request for a PID.	Source: HOPE
Resolver	A piece of software that is able to receive the PID of a resource and to return associated data in a defined form, such as the location of that resource in the form of a URL. In the Handle System, a Handle Proxy Server is a resolver.	Source: HOPE and Australian National Data Service
Reverse Look Up	The procedure which retrieves a PID for an item through a search of its associated data. In the HOPE PID Service, the local identifier can be used to perform reverse look up for the PID.	Source: HOPE See also: *Local Identifier*

4. Digital Objects

Term	Meaning	Comment
Born-Digital Original	The born-digital object in its original quality version, from which all other versions or *Derivatives* can be derived.	Source: HOPE
Compound Object	*Digital Object* composed of multiple content files, for example a periodical issue composed of 25 TIFF files. Structural Metadata describe the internal structure of Compound Objects.	Source: PREMIS (with editing) See also: *Digital Object* *Structural Metadata*
Derivative	Different versions derived from the *Master* or from the *Born-Digital Original*. Derivatives are generally used for web access to digital content. Derivatives may include thumbnail, preview, high- and low-resolution, and OCRed text versions. In HOPE, “Derivative” can be a qualifier; we speak of a “Derivative File” or “Derivative Object” as applicable.	Source: HOPE See also: *Derivative 1 (2, 3)*
Derivative 1	High-resolution *Derivative* for reproduction and publication (online/print) purposes.	Source: HOPE
Derivative 2	Medium to low-resolution *Derivative* for online consultation (view/listen) purposes.	Source: HOPE

Derivative 3	Preview-quality *Derivative* (lowest resolution) for display purposes in search results.	Source: HOPE See also: *Preview* *Thumbnail*
Digital File	Named and ordered sequence of bytes that is known by an operating system. A File can be zero or more bytes and has access permissions, a format, and file system statistics, such as size and last modification.	Source: PREMIS (with editing) See also: *Master* *Derivative*
Digital Object	Discrete unit of information in digital form. In HOPE, a Digital Object instantiates or embodies an intellectual entity, such as a book, a periodical issue, an archival document, a photograph or an audio-visual recording, etc. These objects may be digitised or born-digital. One Digital Object may consist of many *Digital Files*. For example, a digitised book of 300 pages may be a Digital Object consisting of at least 300 files.	Source: PREMIS (see: “Representation”) See also: *Compound Object* *Simple Object*
Europeana Portal Image	A *Derivative 2* standard setting meeting stated file specifications recommended for submission to Europeana. Europeana uses the image to generate four distinct Europeana *Previews*. For more details, refer to Europeana Portal Image Policy.	Source: HOPE
Master	Result of digitisation process: a high-quality *Digital Object* or *Digital File* from which all other versions or *Derivatives* (e.g. compressed versions for accessing via the Web) can be derived. The Master is usually created at the highest suitable resolution and bit depth that is both affordable and practical. In HOPE, “Master” is generally used as a qualifier; we speak of a “Master File” or “Master Object” as applicable.	Source: HOPE See also: *Derivative* *Digital File*
Preview	A *Derivative 3* used for search purposes, showing only a small part of the original. Previews include Thumbnails (of images) and Stills (of films). Default icons for each material type are used when no derivative 3 can be provided.	Source: Glossary v.1 See also: *Derivative 3* *Thumbnail*
Simple Object	*Digital Object* composed of a single content file, for example a report composed of a single PDF file, a manuscript composed of a single JPEG file, or a radio broadcast	Source: PREMIS (with editing)

	composed of a single WAV file.	See also: *Digital Object*
SOR Derivative Table	A table which presents an overview of the derivatives generated by the SOR based on and depending on the quality and format of the master file submitted	Source: HOPE Use for: *SOR File Format Table*
SOR File Format Table	Use: *SOR Derivative Table*	
SOR Processing Instruction	The XML format for exchange of metadata about digital objects between the SOR and the CP.	Source: HOPE
Thumbnail	A preview-quality *Derivative*, used in *Discovery Services* for the discovery, identification, and selection of visual items within collections.	Source: HOPE See also: *Derivative 3* *Preview*

5. IPR

Term	Meaning	Comment
Access Rights	The information that identifies the legal access restrictions pertaining to the *HOPE Social History Resource*, relating to legal frameworks such as the Copyright Laws, Privacy Law, etc. and licensing agreements between CPs and rights owners.	Source: HOPE See also: *HOPE Access Conditions Matrix*
Copyright	Copyright is a set of exclusive rights granted to the author or creator of an original work, including the right to copy, distribute and adapt the work. Copyright owners have the exclusive statutory right to exercise control over copying and other exploitation of the works for a specific period of time, after which the work is said to enter the public domain.	Source: Wikipedia Public domain: see also *Public Content*
Creative Commons License(s)	Creative Commons licenses are several copyright licenses that allow the distribution of copyrighted work	Source: Wikipedia Note that besides licenses, Creative Commons also offers a way to release material into the public domain (*Public Content*) through CC0, a legal tool for waiving as many rights as legally possible, worldwide
Donor Restrictions	A limitation placed on access to or use of materials that has been stipulated by the	Source: SAA Note that donor restrictions

	individual or organization that donated the materials (in practice: donated to the *HOPE Content Provider (CP)*	may require that the collection, or portions of the collection, be closed for a period of time (also referred to as: embargo).
IPR	Intellectual Property Rights are like any other property rights – they allow the creator, or owner, of a patent, trademark, or *Copyright* to benefit from his or her own work or investment.	Source: WIPO
Public Domain	Content in public domain: *Digital Objects* that are publicly accessible without any restrictions (no copyright, no payment restrictions). Whether content is public or not is a matter of *IPR* but also a matter of local content provider policy decision (in terms of pricing policies). In HOPE we speak of content in public domain if direct access to the content, without any restrictions or conditions, is enabled.	Source: Glossary v.1 Note that Europeana has joined with Creative Commons in developing Europeana's Usage Guide for public domain works, which is associated with the Creative Commons Public Domain Mark.

6. Dissemination

Term	Meaning	Comment
Content Profile	Use: *Dissemination Profile*	
CP Dissemination Profile	HOPE CP-specific *Dissemination Profile* for a specific *Discovery Service*, which overrules the HOPE default profile for that discovery service.	Source: HOPE See also: *Dissemination Profile* *HOPE Dissemination Profiles*
Dissemination Profile	A Dissemination Profile provides the rules according to which the *HOPE Aggregator* can select the set of metadata records that need to be disseminated to a given *Discovery Service*. For each discovery service there is one Dissemination Profile. A Dissemination Profile specifies the rules on the basis of metadata values (e.g. IF access condition = open AND link to digital object is available THEN disseminate to Europeana).	Source: HOPE Use for: *Content Profile* See also: *CP Dissemination Profile* *HOPE Dissemination Profile*
HOPE Access Conditions Matrix	The look-up table defining HOPE specific access conditions to *Digital Objects* along three dimensions: content dissemination format (see also *Derivative*), intended use	Source: HOPE

	(fair use/publication), and imposed restrictions (according to agreements between CPs and rights owners).	
HOPE Delivery API	The fully automated method in which *Digital Objects* can be technically accessed and retrieved from the SOR.	Source: HOPE
HOPE Dissemination Profile	This is the set of HOPE default *Dissemination Profiles* for Europeana, the IALHI Portal, and the social sites. Note: Dissemination Profiles are neither *Collection* based nor *Data Set* based.	Source: HOPE See also: *Dissemination Profile*