This is the accepted version of the paper published as:
Bardi, A., Manghi, P., & Zoppi, F. (2014). Coping with interoperability and sustainability in cultural heritage aggregative data infrastructures. In International Journal of Metadata, Semantics and Ontologies (Vol. 9, Issue 2, p. 138). Inderscience Publishers. https://doi.org/10.1504/ijmso.2014.060341

Coping with Interoperability and Sustainability in Cultural Heritage Aggregative Data Infrastructures

Alessia Bardi, Paolo Manghi* and Franco Zoppi

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo",
Consiglio Nazionale delle Ricerche, Pisa, Italy,
Fax: +39 050 315-3464,
E-mail: name.surname@isti.cnr.it
*Corresponding author

Abstract:
The Cultural Heritage (CH) community is one of the most active in the realisation of aggregative data infrastructures (ADIs). ADIs provide tools to integrate data sources to form uniform and richer information spaces. The realisation of ADIs for CH must be based on technology capable of coping with complex interoperability issues and sustainability issues. In this paper, we present the D-NET Software Toolkit framework and services, devised for the realisation of sustainable and customisable ADIs. In particular, we demonstrate the effectiveness of D-NET in the CH scenario by describing its usage in the realisation of a real-case ADI for the EC project Heritage of the People's Europe (HOPE). The HOPE ADI uses D-NET to implement a two-phase metadata conversion methodology that addresses data interoperability issues while facilitating sustainability by encouraging participation of data sources.

Keywords: cultural heritage; data infrastructures; D-NET; metadata; aggregation; metadata transformations; service-oriented

Biographical notes:
Alessia Bardi received her MSc in Information Technologies in the year 2009 at the University of Pisa, Italy. She is a PhD student in Information Engineering at the Engineering Ph.D. School "Leonardo da Vinci" of the University of Pisa and works as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. Her research interests include Digital Library Management Systems, data interoperability, compound object management, and data infrastructures for e-science and scholarly communication.

Paolo Manghi received his PhD in the year 2002 from the Dipartimento di Informatica of the University of Pisa, Italy. He is presently working as a researcher at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. His research interests include Data Models for Digital Library Management Systems, Types for Compound Objects, data curation in Digital Libraries, service-oriented ICT infrastructures with special focus on data ICT infrastructures.

Franco Zoppi has been working since the 1980s on the design and implementation of software systems in the areas of DBMS, Distributed Office Information Systems and Digital Library Systems. Initially employed at the Research and Development Department of Olivetti S.p.A., then at the Network Laboratory of the Telecommunications Department of Telecom Italia, in 2001 he joined the Information System Department of Pisa University as Project Manager. Since 2005 he has been working as Research Associate at CNR-ISTI, where he coordinates the CNR activities in several EU-funded projects.

1 Introduction

In the last decade, the multi-disciplinary character of science and the need of researchers to gain immediate access to research material often led to the realisation of so-called *aggregative data infrastructures* (ADIs). These are here intended as information systems where organisations (e.g. research centres, universities, industries) can find the tools to integrate their data sources to form uniform and richer information spaces and support their communities with enhanced access services to such content. In particular, ADIs offer functionality for (*i*) the collection and processing of metadata descriptions of files (digital objects) in order to populate a uniform aggregated information space and (*ii*) the provision of the information space to

humans, via web portals, and machines, via standard APIs. On the one hand, one major challenge for ADI designers and developers is to provide tools capable of dealing with several interoperability issues derived by the mismatch between the aggregated information space and the data sources; e.g. export protocols, structure and semantics of metadata, physical representation. On the other hand, another grand challenge is to realise ADIs capable of coping with the dynamic and complex requirements of research communities, whose needs in terms of content, functionality, and quality of service tend to vary along time, as science evolves. Indeed, software and system refinements prove to be as expensive as necessary for the ADI to grow and be up to the challenge of its community. Therefore, the adoption of the proper enabling technology plays a crucial role for the sustainability of an ADI. Such technology should minimise the cost of design and development required to realise, operate, and modify data infrastructures.

The *D-NET Software Toolkit* (Manghi et al., 2010c) (D-NET Lab, 2009) was specifically realised to facilitate designers and developers in the construction and maintenance of ADIs. D-NET implements an open source service-oriented framework where services for the collection, processing and provision of metadata and files from a set of data sources can be customised and combined to implement the internal workflows of ADIs. As proven by the several installations and adoption in a number of European projects – DRIVER (DRIVER, 2007), DRIVER II, OpenAIRE, OpenAIREplus (OpenAIRE, 2010), EFG, EFG1914 (Artini et al., 2013) (EFG, 2010), ESPAS (ESPAS, 2012) – ADIs realised with D-NET are easily customisable, extensible, scalable, and sustainable (Manghi et al., 2010a).

In this work we focus on the Cultural Heritage (CH) domain, which is certainly one of the most active in the operation of ADIs (Blanke, 2010) (Wang et al., 2012). The increased availability of CH digital content raised a natural need to deliver ADIs for the integration and delivery of such content to wider research, academic, and public communities (Cucchiara et al., 2012) (Loebbecke and Thaller, 2011); examples are the ADIs supported by Europeana (Europeana Foundation, 2009) and its satellite projects. Facilitating content interoperability, which naturally surfaces in a social environment (Alemu et al., 2012), is certainly one of the main challenges in CH, e.g. CIDOC approach (Doerr, 2003) (Stasinopoulou et al., 2007), and in constructing ADIs. In particular, the realisation of ADIs for CH carries a higher level complexity on this respect when compared to other disciplines. The reason is the high degree of heterogeneity brought in by CH communities (Papatheodorou, 2012) (McDonough, 2008), which are typically formed by groups of sub-communities whose research focuses may diverge but require to be connected to enable better science. Their metadata is not only syntactically and semantically heterogeneous, but multilingual, semantically rich, and highly interconnected.

In this paper, we show how the D-NET Software Toolkit can be an ideal candidate to satisfy the interoperability and sustainability requirements of ADIs for CH. Specifically, we show how its services can be used to implement a *two-phase metadata conversion methodology* which softens the interoperability issues arising in CH scenarios featuring highly heterogeneous data sources. To this aim, Section 2 presents the evolving requirements surfacing when realising ADIs and the sustainability issues they entail for supporting organisations. Section 3 describes the D-NET Software Toolkit framework and services for the construction of ADIs for CH. Finally, in Section 4 we describe how D-NET is used to instantiate a two-phase conversion ADI in the context of The Heritage of the People's Europe (HOPE) project (HOPE, 2011). The objective of the HOPE ADI is to offer a unique entry point to digital objects for the social and labour history from the 18th to the 21st century in Europe. It federates digital object collections from several major European institutions in the field. HOPE is an exceptionally representative scenario of CH's richness, since social and labour history covers a wide range of digital objects, such as documentaries, pictures, drawings, and archival documents, in turn described by highly heterogeneous metadata representations.

2 Aggregative Data Infrastructures

In the last few years, an increasing number of research communities started federating their data sources into ADIs. A high-level functional architecture of ADIs is shown in Figure 1: ADIs are intended here as systems capable of collecting *metadata records* and *files* relative to objects stored in a set of heterogeneous *data sources*, in order to construct an homogeneous information space of metadata records conforming to a *common data model* (Stasinopoulou et al., 2007). On top of the resulting information space, ADIs provide community services that support advanced access to the aggregated data; e.g. cross-source search and browse, cross-source object interlinking, standard API exports, etc. ADIs typically focus on metadata aggregation and realise information spaces whose data can be used to cross-search over files which are kept at their original locations. In some scenarios, the files may be collected or uploaded in an ADI, which may offer services for digital preservation (Cramer and Kott, 2010; Li and Banach, 2011) or services for feature/information extraction.

In the following we shall describe the functional areas of ADIs and the two main challenges to be tackled when realising such systems: data interoperability and curation, and coping with evolving requirements.

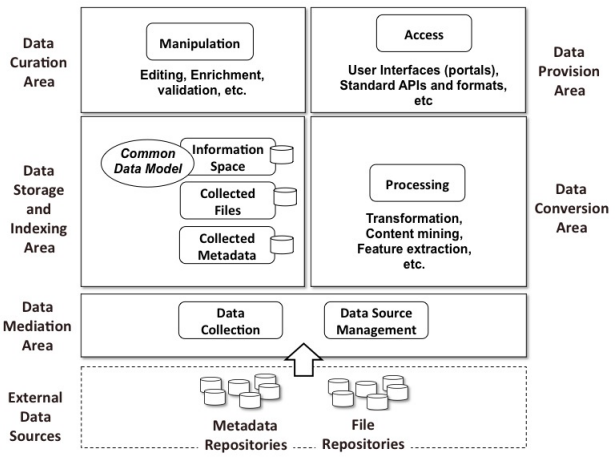

Figure 1 Aggregative Data Infrastructures

2.1 Functional areas of ADIs

The high level architecture in Figure 1 shows how ADIs can be conceived as a set of components organised in five functional areas:

Data mediation area. Components in this area are capable of managing a set of external data sources and of collecting metadata records and files compliant to different data models and formats.

Data storage and indexing area. Components in this area manage storage and access to collection of objects conforming to a given data model. Examples are: an index component handling Dublin Core metadata objects (Dublin Core Metadata Initiative, 2008), a storage component for the preservation of large collections of video files or metadata records.

Data conversion area. Components in this area offer functionality for processing metadata records and files. Examples are: transformation from one data model into another (e.g., from DOC payload objects to PDF, from MARC (Library of Congress, 2005) metadata records to Dublin Core's), extraction of information from objects (e.g., extracting histograms from image payloads, extracting full-text from PDF files, inferring language from text).

Data curation area. Components in this area provide data curators, i.e. administrators of the ADI, with tools to enrich, update, insert and delete objects of the information space, exploiting feedback from data source managers. This area may include components for the evaluation assessment and monitoring of the aggregative information space (Fenlon et al., 2012), providing data curators with feedbacks to improve conversion workflows. Moreover, components for the validation of objects w.r.t. content quality policies can also be provided, to serve data source managers at improving their local collections.

Data provision area. Components in this area allow third-party applications to access objects in the storage area according to standard APIs; e.g., OAI-PMH (Carl Lagoze and Herbert Van de Sompel, 2003), OAI-ORE (Lagoze and Van de Sompel, 2006), FTP, WSDL/SOAP, SRW, REST.

2.2 Data interoperability and curation

ADIs collect from data sources, through standard APIs, files and relative metadata records. In the following we shall focus on XML metadata data sources, that is data sources exporting their content as XML metadata records in XML format, but similar reasonings can be applied when other formats are available. For example RDF in its various manifestations, from OAI-ORE (Hunter and Lagoze, 2001) to the new trend of LinkedData in CH (Hyvönen, 2012) (Haslhofer et al., 2010), (Isaac et al., 2012). Metadata records are on-the-wire representations of data conforming to the data source data model. The ADI information space contains data conforming to the given common data model whose physical representation may be based on several standard storage solutions, such as relational databases, graph stores, full-text indices, XML native stores, etc. ADIs must therefore provide tools to overcome two main interoperability barriers: the definition of structural/semantic mappings from data source data models onto ADI common data model and the definition of physical mappings from XML metadata records to ADI storage data representation. Specifically, the design and implementation of ADIs must face the following technical interoperability challenges (Manghi et al., 2010a):

Mediation interoperability. Data sources may export metadata and files according to different standard protocols. Typically, ADIs solve this issue by natively supporting standard exchange protocols, such as OAI-PMH, FTP, HTTP, and including services capable of collecting and storing data locally.

Representation interoperability. As mentioned above, collected metadata records are encoded in XML while data in the ADI information space may not necessarily be stored in the same way. Conversion software must therefore encode both logical and physical mappings from XML records onto information space objects. Typically, ADIs facilitate this task by defining a common XML schema for representing the information space data model. This leaves the logical mappings at the level of XML schemas, where XSLT mappings can be flexibly and more easily defined for each data source. Physical mappings, i.e. code to transform XML records into information space objects, is written only once. Figure 2 exemplifies the described approach. The data source adopts the MARC data model and delivers metadata records as MARC-XML files. The ADI information space adopts the Dublin Core data

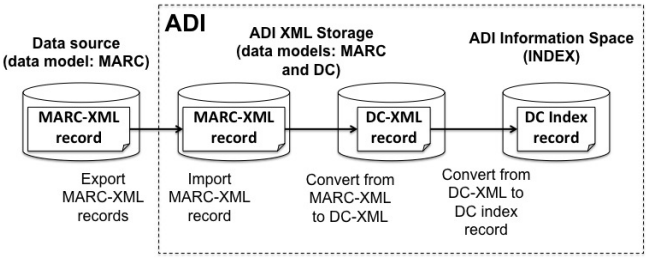

Figure 2 Interoperability of metadata representation

model and represents Dublin Core objects as "indexable Dublin Core documents". In order to aggregate records from the data source, the ADI is configured to collect and transform each MARC-XML record onto a DC-XML record; then, to convert the resulting DC-XML record into its corresponding index representation.

Structural and semantic interoperability of metadata. Collected metadata records are encoded in XML but according to structure (XML schemas) and semantics (e.g. vocabularies, value formats) which differ from data source to data source. Semantics and structure depend on the data source data model, i.e. the entities and relationships used to describe or contextualise the digital objects at hand, but also on the underlying storage platform. Typically, ADIs solve this issue by including services capable of mapping input XML metadata records onto XML records conforming to the common metadata schema. Such mappings (e.g. XSLT scripts, executable code) are implemented by data curators who define structural (e.g. paths to paths) and semantic (e.g. vocabulary terms to vocabulary terms) correspondences inspired by feedback from data source managers. We can distinguish between two main approaches: *format-to-format* mappings, via XSLT or executable code, and *format-to-common ontology* mappings, which indirectly enable inter-format mappings (Gaitanou et al., 2012).

Granularity interoperability of metadata. By *granularity* we mean the level of data model detail represented by one XML metadata record. In some cases each record represents one entity of the model (e.g. a Dublin Core record represents and describes one publication entity), in other cases it may represent more entities possibly with relationships between them. For example, one ESE (Europeana Foundation, 2012) record can represent a set of entities, while an EAD (Library of Congress, 2002) record may represent a hierarchy of entities, as depicted in Figure 3. Since structural and semantic mappings for XML records apply at the level of the individual entities (i.e. one record represents one entity), in such cases further services are required to *unpackage* records in order to single out the XML representation of the entities before applying the mappings.

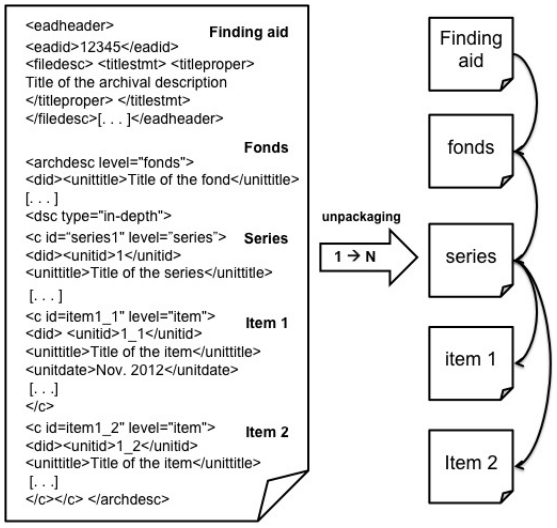

Figure 3 Achieving granularity interoperability

Manipulation of information spaces. The ADI common data model is typically defined to minimise information loss w.r.t. the collected data and relative data models, but also maximise the quality and richness of the generated metadata records (e.g. entity properties should rarely have missing values). In some cases, the model includes attributes whose values may be derived by extracting information from the metadata records (e.g. mining attribute values and relationships between records to infer further values or relationships) or from the files described by such records (e.g. histograms from images, keywords from text documents). Typically, ADIs solve this issues by including services capable of processing the collected data to enrich the quality of the records in the information space. Moreover, collecting objects from several data sources may lead to duplication of content, whenever different sources keep information about the same entities. In such cases, de-duplication actions, i.e. merge metadata records describing the same object into one, may be necessary to disambiguate the information space. To this regard, ADIs may include de-duplication services specifically devised to exploit attributes and relationships of a record to identify and (semi-)automatically merge similar records.

2.3 Evolving requirements

Organisations willing to realise ADIs must be able to sustain the initial design and development cost, plus the refinement costs made necessary by further changes required by the operative ADIs. Indeed, ADIs are often characterised by highly evolving requirements in terms of content, functionality, and Quality of Service (QoS). On the content side examples are changes to the common metadata model, new mappings required to handle interoperability with new joining data sources, etc. On the functionality side examples are changes in the data management workflows, new services to integrate missing functionality, etc. For example, a workflow for

collection, conversion, storage, and indexing may turn into a workflow for collection, storage, conversion, and indexing workflow, to make the index more efficiently re-generated when changing the mappings. On the QoS side, management of storage and index replicas may be required to ensure robustness and availability.

Most ADI enabling software in the literature are designed to tackle very precise data aggregation scenarios and can hardly be re-used in different contexts and domains, examples are the projects Multimatch (Amato et al., 2009), KEEP (KEEP Consortium, 2009), MICHAEL (MICHAEL Culture Association, 2005), DARIAH (Blanke and Hedges, 2008) and CLARIN (Váradi et al., 2008). This is due to the overall absence of general purpose software for ADIs, which leads organisations to face the high cost for the realisation of their ADIs from scratch and in a very pragmatic way. The typical approach is the integration of existing open source technologies and products, such as OAI-PMH aggregators (DLXS (University of Michigan, 2006), Repox (Reis et al., 2009)), full-text indices (Apache Lucene and Solr (Apache.org, 2011)), XML stores (e.g. eXistdb (exist Solutions, 2012), Sedna (Institute for System Programming RAS, 2011)), etc. As a consequence, the result are ADIs which very efficiently address their initial requirements, but involve high refinement and maintenance costs whenever the dynamic requirements described above must be satisfied. In many cases, organisations must face the trade-off between refinement costs and end-user satisfaction.

2.4 Software systems for the realisation of ADIs

The realisation of ADIs is not trivial in terms of technical expertise, development and maintenance costs. The traditional approach of creating "from scratch" such infrastructures turns out to be not affordable in the majority of cases (Manghi et al., 2010c). Typically, developers re-use existing software products and code interoperability layers to implement their integration into an ADI. Such solutions are hardly re-usable in different ADI contexts and are in general not sustainable in terms of software maintenance (e.g. lack of documentation and continuity) and integration of new functionalities (Manghi et al., 2010a). As a reaction to such drawbacks, research in the e-Infrastructure field started investigations on software systems specifically designed to support the creation of ADIs (Candela et al., 2013; Manghi et al., 2010a). Typically, such software systems implement modules supporting general-purpose functional patterns for data collection, processing, storage and provision in order to allow developers to build ADIs by re-using, customising, and pipe-lining functionalities into workflows to meet the specific community needs. Examples of such systems, focusing on metadata collection are SYNAT (Mazurek et al., 2013) (Rosiek et al., 2013), CORE (Knoth and Zdrahal, 2012), and MoRe, i.e. the Monument Repository (Dimitris Gavrilis, 2013). SYNAT and CORE offer advanced and configurable services for the construction of ADIs for the scholarly communication (Castelli et al., 2013), i.e. aggregation and curation of metadata collected from heterogeneous publication repositories. As such, they would be too limited to realise the expected ADI features. MoRe was devised within the CARARE project (Papatheodorou et al., 2011), i.e. an aggregator of archaeological data providers for Europeana, in order to offer ADI construction capabilities for archaeological content, e.g. sites and monuments information. To this aim, MoRe implements (micro-)services for metadata ingestion, mapping, OAIS-compliant preservation, geo-data curation, quality monitoring and semantic enrichment. Services can be flexibly customised and combined into autonomic workflows, but their functionalities are tailored to the realisation of ADIs for archelogical archives.

To our knowledge, the literature does not report on "enabling software" for the construction of ADIs for CH, but rather on ad-hoc software solutions to serve specific ADI use-cases or classes of ADIs in this domain. The D-NET Software Toolkit, described in the following section, was first developed to address use cases similar to those targeted by SYNAT and CORE. The first ADIs based on D-NET was realised in 2007 for the EC-founded project DRIVER (Digital Repository Infrastructure Vision for Europe) (Feijen et al., 2007), whose aim was to aggregate from institutional repositories all across Europe, metadata about scientific articles. Since then, D-NET has been extended and evolved in order to satisfy the requirements of other types of community, such as those in the cultural heritage and scientific data.

3 The D-NET Software Toolkit

In this section we present the D-NET Software Toolkit (D-NET Lab, 2009)(Manghi et al., 2010b) and show how it covers the technical challenges described in Section 2. D-NET is an open source, general-purpose software conceived to enable the construction and operation of ADIs, based on given initial requirements, and to facilitate their evolution along time to satisfy future requirements. D-NET implements a service-oriented framework based on standards, namely Web Services with SOAP and REST APIs, where ADIs can be constructed in a LEGO-like approach, by selecting, customising, and properly combining D-NET services. The resulting ADIs are run-time systems which can be flexibly re-configured and extended (e.g. new services can be integrated), and can scale in terms of storage and workload (e.g. storage and index replicas can be maintained and deployed on remote nodes to tackle multiple concurrent accesses or very-large data size).

D-NET offers a rich and expandable set of services targeting data collection, processing, storage, indexing, curation and provision aspects. Services are defined to be modular, configurable (e.g. by data model), and designed to be included as steps of configurable pipelines i.e.

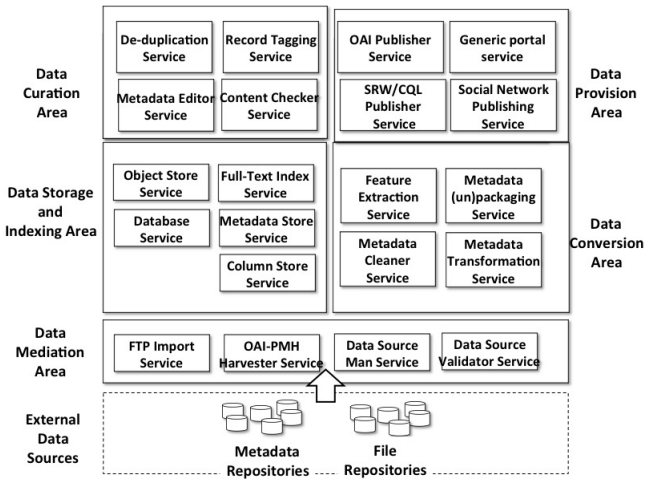

Figure 4 The D-NET services

they implement common data exchange interfaces. As such they can be customised and combined to meet the data workflow requirements of a target user community. Moreover, they can be partly or fully replicated and distributed over different servers depending on the QoS needs of the specific community. In general, multiple instances of a service improve fault tolerance, reduce the overload of each instance, and make it possible to dynamically re-organise the environment when a server is not reachable. Figure 4, illustrates the D-NET data management services, some of which realised in the context of the HOPE project, and how these implement the high-level architecture and functionalities of ADIs.

Data Mediation Area. Services in this area are capable of managing (register and de-register to the ADI) a set of available external data sources and of collecting their objects. D-NET offers services for on-demand and programmatic data collection based on the following standard protocols: OAI-PMH, FTP, FTPS (FTP over SSL/TSL), SFTP (SSH File Transfer Protocol), HTTP/HTTPS.

Data Storage and Indexing Area. Services in this area manage storage and access for files and metadata records. Services offer various data storage supports, abstracting over relational databases (Postgres), file storage (MongoDB (MongoDB, 2012) or standard file system), full-text indices (Apache Solr (Apache.org, 2011)), column stores (Apache HBASE (The Apache Software Foundation, 2013)), and metadata store (abstraction on top of file storage services). Developers can configure and choose the most proper storage based on the functional requirements and the common data model of the ADI at hand.

Data Conversion Area. Services in this area offer functionalities to convert input XML metadata records, regardless of the structure and semantics of their schemas, and files, regardless of their storage formats, and produce list of output XML records.

The Transformation Service can be configured to transform metadata records from one schema to another (e.g. from MARC to Dublin Core) given XSLT mappings. D-NET data managers can create, update, and remove such mappings and configure the service to apply one mapping to a given input (e.g. a metadata store) at given time intervals. Mappings are therefore persisted and available for re-use by other data managers and across different crosswalks. Note that data managers, possibly supported by experts in the fields, are also in charge of verifying the quality/effectiveness of the mappings and eventually of their refinement. In particular, in the case of one-to-one mappings between XML records, D-NET provides Transformation Services with end-user interfaces for the aided-creation of mappings in the style of Repox (Reis et al., 2009) and other similar tools – the service was developed at the University of Bielefeld (ref. Jochen Shirrwagen and Friedrich Summann). In general, transformation mappings between different formats can be transitively combined to obtain indirect metadata conversions.

The Metadata Cleaner Service harmonizes values in metadata records based on a set of thesauri. A D-NET thesaurus consists of a *vocabulary* that is a list of authoritative *terms* together with associations between terms and their *synonyms*. Data curators – typically based on instructions from data providers and domain experts – are provided with user interfaces to create/remove vocabularies and edit them to add/remove new terms and their synonyms. Given a metadata format, the Metadata Cleaner Service can be configured to associate the metadata fields to specific vocabularies. The Service, provided records conforming to the metadata format, processes the records to clean field values according to the given associations between fields and vocabularies. Specifically, field values are replaced by a vocabulary term only if the value falls in the synonym list for the term. If no match is found, the field is marked as "invalid". The "invalid" mark is exploited by the Content Checker Service (see Data Curation Area) to help data manager at refining the thesauri or data providers at improving the quality of their data.

The metadata record conversion suite is completed by (Un)Packaging Services which solve granularity issues by (un)packaging XML records (one-to-many and many-to-one conversions) based on a number of (un)packaging modules (business logic) to be made available by data managers. Such modules go beyond the limits of XSLT and can integrate Java business logic.

Finally, the Feature Extraction Service can perform information extraction from files, be them XML records or other formats, according to given algorithms. The result is a list of corresponding XML records containing the information inferred from each file, to be used for application specific purposes. For example, extraction algorithms may infer the language of the text in PDF

files; the languages resulting from batch-processing a list of PDFs could be set as input of the Packaging Service together with the related metadata records, in order to create a version of the descriptions inclusive of the inferred languages.

Data Curation Area. Services in this area offer functionalities for data curation and enrichment. The Content Checker Service provides data curators with an overview of the information space, where they can search ad browse records and verify the correctness of the conversion phase (e.g. no mapping mistakes or semantic inconsistencies, no records marked as "invalid" by the Metadata Cleaner Service). Upon positive verification of the records in the Information Space, data curators can mark the content from a given data source as visible to the public.

The Metadata Editor Service allows data curators to add, edit and delete metadata records once they have been aggregated in the information space. The service takes as input the schema of the information space, which describes record properties as well as relationships between records, and respecting its structure enables edit actions such as changing record property values or creating relationships between records. Data curators are supposed to be experts in the application domain at hand. They are associated to the edit actions they apply, hence fully responsible for the quality and authoritativeness of such updates. The service acts as a "record patcher", meaning that changes are persisted independently from the edited records to be applied to the last available version of the records (e.g. last harvested and transformed) before these are streamed to the next step in a workflow. This particular need derives from the aggregative scenario, which requires frequent "data refresh" operations, i.e. re-collection of data from original data sources. Since original records may change from different refreshes, the service ensures that the last version of the data is always available and that data curators updates are applied on top of that.

The De-duplication Service (Manghi and Mikulicic, 2011) allows data curators to disambiguate (BigData) information spaces by merging duplicate records and operates with parallel Map Reduce jobs over the Column Store Service. The service identifies and returns the groups of records candidate for merging – i.e. equivalent records, in the sense they represent the same real-world entity – based on sorted neighbourhood algorithm with blocking and a record similarity function that is configurable by data curators. It analyses a collection of records with identical structure, namely a flat list of repeatable fields, based on a configuration of blocking and similarity functions, and a similarity measure threshold (i.e. $0 \dots 1$). Typically, the De-duplication Service is fed with surrogates of the information space records that consist of the identifier and the record properties suitable to calculate a similarity distance. The actual merge of similar records in the information space is left to software developed by ADI developers, since the strategy cannot be captured by generic templates.

D-NET was extended in HOPE to include the Record Tagging Service, which allows data curators to bulk-tag a group of objects in the information space with the purpose of object categorization. The service can be configured to include user-defined classification schemes and relative categories, identified by tags. It provides data curators with a virtual environment where they can (i) search and browse to identify the sets of objects they believe should belong or not belong to a given category, identified by a tag, (ii) eventually perform the tagging and untagging actions required to assign or remove the intended category to such a set, and (iii) preview the effects of their actions before making the changes visible to the end-users.

Finally, D-NET services for measuring metadata quality w.r.t. given guidelines are available but only for the Dublin Core format (developed by the Madgik Team, Computer Science Department of the University of Athens). The extension of such services to cover arbitrary metadata formats is in the plan.

Data Provision Area. Services in this area allow third-party applications to access objects via standard APIs. D-NET currently supports the following provision protocols: OAI-PMH (enabling harvester to access metadata records), SRW, REST and WSDL/SOAP (enabling third party applications, such as portals, to perform queries on D-NET indices).

User Interface Services can be used to automatically generate templates of portals based on a given XML schema, e.g. the ADI common data model, used in one of the Index Services of the ADI.

Finally, in the context of HOPE D-NET was extended with Social Network Publishing Services for the automatic export of videos and pictures towards social networks. Currently, two exporting modules are available for YouTube and Flickr, but others can be added in the future. In HOPE the service is used to publish files whose metadata records bear a special tag, to be assigned by data curators via the Record Tagging Service.

4 D-NET in the Cultural Heritage

UNESCO's definition of cultural heritage (UNESCO, 1972) lists a vast number of types of objects preserved and represented in the archives, libraries and museums: some examples are paintings, buildings, landscapes, pictures, posters, sculptures, documents. Integrating and interlinking such objects represents a field of research and offers new opportunities of investigation to scientists.

Social history is an important field of the CH domain. Most of the social history collections are about movements and people who opposed the state, capitalism and the established order, such as trade unions, left-wing political parties, revolutionaries, anarchists, but

also environmental activists, women's rights movements. Those activists were persecuted and sometimes they were forced to leave their countries. Therefore, archives and libraries with materials about the opposition to the state were often smuggled out to foreign countries, in order to keep them safe from confiscation and destruction. As a consequence, a lot of social history material is today preserved by small, independent archives, libraries and research institutions all across Europe, rather than by national repositories.

The virtual integration of those scattered collections is very attractive for historians, who usually need to consult tens of different and geographically distributed archives and libraries to find materials about their research topic. Historians could benefit from an integrated CH information space where it is possible to consult content of hundreds of archives and libraries in one click, even accessing to small repositories which were not previously known by the researcher. For example, Figure 5 shows a subset of the results of a search in the Social History Portal (HOPE, 2013) realized in the context of the HOPE project. The search is about "August Bebel", one of the founders of the German Social Democratic Party, who lived from 1840 until 1913. The search returns around 300 documents including images and texts from several different European institutions. In Figure5 we report some results from the AMSAB Institute of Social History in Belgium, the International Institute of Social History in Netherlands, and the Archiv der Sozialen Demokratie in Germany. Figure 6 shows instead a collage of the tools offered by the same portal to allow browsing through the aggregated objects via a time-view and a map view, filtering by theme, country, provider, and language.

The Social History Portal accesses the HOPE information space by exploiting functionalities offered by the HOPE ADI. Goals and challenges of the HOPE project are presented in Section 4.1. Section 4.2 describes how the HOPE ADI addresses those challenges and implement the functional requirements of the HOPE community. Details about the approach used to solve data interoperability issues are given in Section 4.3. Finally Section 4.4 summarises the achievements of the HOPE project.

4.1 The HOPE Project: Overview and Challenges

HOPE (Heritage of the People's Europe, FP7 EU eContentplus, grant agreement: 250549) (HOPE, 2011) is a "Best Practice Network" for archives, libraries, museums and institutions operating in the fields of social and union history. The goal of the project is providing a unified access to materials about the European social and labour history from the 18th to 21st centuries, proposing guidelines and tools for the management, aggregation, harmonisation, curation and provision of digital CH content. Institutions, i.e. *content providers*, joining the HOPE network benefit of an advanced, distributed ADI. The ADI enables them to enhance the

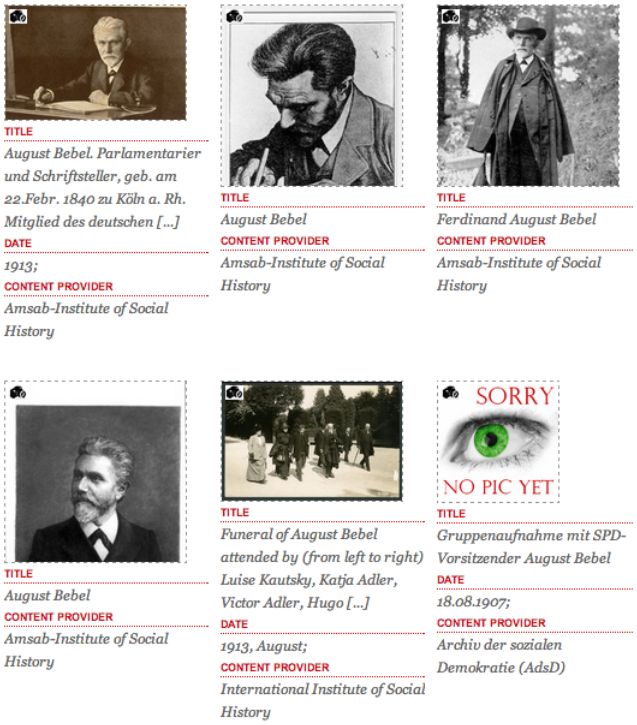

Figure 5 Searching in the Social History Portal: material preserved by different institutions can be accessed with one single click

quality and the visibility of the digital cultural objects preserved in their *data sources*. At the end of the project, HOPE will collect metadata records describing around 2,000,000 resources in the CH domain.

The functional requirements of the HOPE community can be summarised as follows:

- The HOPE ADIs must be able to handle a dynamic number of content providers and relative data sources, each delivering XML metadata records of different formats (i.e. obeying to different XML schema) and via different export protocols (e.g. OAI-PMH, FTP, HTTP). Figure 1 summarizes the heterogeneity of the input data sources.

- The HOPE ADI must minimize the loss of quality of the aggregated metadata records. As a consequence, the ADI must adopt a rich data model that include fields specific to the CH domain.

- The HOPE ADI must promote the enrichment of the aggregated metadata records by providing data curators with tools for metadata curation, i.e. harmonisation and tagging, and metadata quality control, i.e. error detection and content checking.

- The aggregated records must be accessible by third parties via programmatic interfaces in order to enable the realization of advanced end-user tools, such as web portals. HOPE records must also be

Figure 6 Searching in the Social History Portal: material preserved by different institutions can be accessed via a time map or a geographical visualization

exported via OAI-PMH in order to be collected by Europeana.

- The HOPE ADI must be able to publish digital objects and relative metadata records to social sites. Each provider must be able to select which digital objects (among those the provider delivered to the HOPE ADI) can be published on which social site.
- The HOPE community wants to improve the accessibility of the digital objects and relative metadata records by providing persistent identifiers where these are missing.
- HOPE partners must be provided with a repository system they can use to store their digital objects in case they cannot afford the scalability cost of a local object file store.

In the next section, we shall describe how the HOPE ADI implements the above requirements by means of D-NET services. An exception holds for the last two requirements for persistent identifiers and digital object management, which were dealt with by services developed at IISG in Amsterdam, The Netherlands. The services are not currently part of the D-NET, although they could be easily integrated in the framework in the future.

The Persistent Identifier service deals with minting and maintenance of persistent identifiers. The service is based on the Handle System (International DOI Foundation, 2012) and is used by either content providers or the ADI to assign PIDs to metadata records and the digital objects they describe. In the first case, content providers use the service to replace their local identifiers with PIDs before they export content to the ADI; in the second case it is the ADI which looks after local identifiers as provided by the data sources not falling in the former category and replaces them with PIDs.

Finally, the so-called Shared Object Repository (SOR) deals with the management (storage, access, and conversion) of digital files. HOPE content providers can deposit their object files into SOR, which automatically applies conversion algorithms to create files in standard formats and with sizes suitable for web dissemination.

4.2 The HOPE Aggregative Data Infrastructure

The HOPE ADI is implemented using and extending with new services the D-NET Software Toolkit. In the following we shall describe how the D-NET software is today used to satisfy the functional requirements described in the previous section in an efficient and sustainable way.

Content providers and data sources. The ADI should be able to handle a varying number of *content providers*, which may be in turn deliver several *data sources*, each dedicated to storage of metadata records and files relative to different object typologies; e.g. an institution may offer two data sources, relative to an archive and a library. Indeed, as it often happens in the CH domain, content providers may deliver data sources whose objects belong to diverse sub-communities (in HOPE's ADI referred to as *profiles*), which in HOPE are: *library*, *archive*, *visual*, *audio video*. Although a profile marks a data source as including material of the same "semantic domain", distinct data sources of the same profile may store objects of different formats (e.g. images, videos, audio, text material) and described by different data models and relative metadata formats. For example, librarians may operate data sources containing files of several formats whose descriptions may conform to data models such as Dublin Core or MARC and be exported according to the relative XML metadata formats. In addition, data sources may export their content via any standard protocols, such as OAI-PMH, FTP, etc. Table 1 illustrates the heterogeneity of the metadata records delivered by content providers to the HOPE ADI – for more information on the content providers visit `http://www.peoplesheritage.eu/content/partners.htm`. Currently, the ADI registers 14 content providers featuring 128 data sources of different HOPE profiles and exposing records conforming to different export formats. In total, data source records are mapped onto around 2,500,000 records relative to digital objects and descriptive metadata. This number is expected to increase as the ADI will attract more content providers.

Common data model. According to investigations run by CH experts of the HOPE consortium, the CH standard data models would be either too specific (e.g. CIDOC CRM, EAD), i.e. not able capture the diversity of the four domains, or too generic (e.g. MARC, Dublin Core) to deliver a concrete and usable ADI. Based on the four HOPE domain profiles, the HOPE Consortium defined a common metadata model and its corresponding XML schema (Lemmens et al., 2011). In order to capture the commonalities of diverse object domains and formats, the model has been defined by studying the characteristics of the four profiles from the perspective of well-established standard formats in the respective field: MARCXML for libraries, EAD for archives, EN 15907 for audio video, and LIDO for visual.

As depicted in Figure 7, seven classes of entities resulted from this process. The *Descriptive unit* entity is central and represents descriptions of CH objects (e.g. date of creation, type of material, title). Based on the identified profiles, the descriptive unit class has four subclasses containing properties that are peculiar to one specific HOPE domain. Specifically, descriptive properties for the *Archive Unit*, *Library Unit*, *AudioVideo Unit* and *Visual Unit* are obtained respectively from EAD (Library of Congress, 2002), MARC (Library of Congress, 2005), EN 15907 (European Committee for Standardization,

Current Content Providers	Domain Profile	Export Metadata Format	Profile Standard Format	# Data Sources (total = 128)	# HOPE records (total = 2,597,484)	
IISG	visual	MARCXML	LIDO	5	312,779	312,779
Persmuseum	visual	MARCXML	LIDO	1	6,077	6,077
VGA	library	MARCXML	MARCXML	2	29,540	29,540
AMSAB	library	MODS	MARCXML	2	19,458	183,893
	archive		EAD	2	97,120	
	visual		LIDO	1	67,315	
CGIL	archive	EAD	EAD	3	48,054	49,633
	library		MARCXML	1	1579	
Generiques	archive	EAD	EAD	4	8,271	8,271
FES Library	library	idiosyncratic	MARCXML	11	403,025	403,025
FES Archive	visual	idiosyncratic	LIDO	7	367,651	367,651
FMS	archive	idiosyncratic	EAD	23	875,929	875,929
TA	visual	LIDO	LIDO	1	55,959	57,537
	general	oai_dc	oai_dc	1	1,578	
UPIP	general	oai_dc	oai_dc	15	5,016	5,016
CEDIAS	general	oai_dc	oai_dc	4	536	536
OSA	archive	FOXML/METS	EAD	9	213,830	243,927
	library		MARCXML	4	30,097	
SSA	visual	idiosyncratic	LIDO	42	53,670	53,670

Table 1 Summary of content delivered by providers to the HOPE ADI

2010), and LIDO (ICOM International Committee for Documentation, 2010). Cross-domain properties are grouped in the descriptive unit super class. Descriptive units are related with each other via *containment* and *sequential* relationships so that it is possible to represents hierarchies of objects (for example a manuscript with miniatures, where there is a description of the whole book - the container - and a description of each miniatures in it) and sequences of objects (for example the sequence of chapters of a book, where each chapter is represented by one descriptive unit). The *digital resource* entity contains descriptive information about a digital representation of the object (e.g. the picture of one side of a coin, the digitised page of a book), technical information (e.g. an URL to the object), and it has a relationship to the corresponding descriptive unit. Digital resources related to the same descriptive units can express sequential relationships, thus establishing a "reading path". *Agents* are persons having to do with the lifecycle of a Descriptive unit. For example, the relationship *isPublishedBy* can link a descriptive unit to an agent who is the publisher of the described object. *Themes* are thematic headings specific to the fields of social and labour history. Objects of this class of entities (i.e. terms of a vocabulary) have been defined by a group of experts in the social and labour history domain from the HOPE partners and are used to add context to Descriptive units. Similarly, *Places* and *Events* are intended as authoritative entities, while *Concepts* include any valuable term, collected from the data sources, that cannot be mapped into an object of the former classes. Themes, Events, Places, and Concepts are meant to contextualise Descriptive units via relationships whose names embody the semantics of the association. However, since defining common vocabularies and creating structural and cleaning mapping from data sources to the objects (terms) of such classes (vocabularies) has a high complexity, the HOPE consortium decided to concentrate the attention on Themes only. This simplification sped up the aggregation of all data sources for immediate use to the HOPE community and Europeana. The mappings will be extended to cover the entities Place, Event, and Concept in a later stage.

An example of HOPE visual Descriptive unit is shown in Listing 1. Row 6 (*isContainedBy*) contains the handle "http://hdl.handle.net/12345/AAAA", which is a reference to the parent metadata record. This means that the record shown is part of a hierarchy. Rows 8-14 (*isRepresentedBy*) bear information about a digital resource that is related to the current record. Row 22 (*visualUnit*) finally contains profile-dependent information: *technicalAttribute* and *objectName* are

Listing 1 HOPE common metadata format: a sample visual descriptive unit

```
<hopeEntity>
    <persistentID>http://hdl.handle.net/12345/ABCD</persistentID>
    <localID>CD_12_1_1</localID>
    <descriptiveUnit>
        <europeanaType>SOUND</europeanaType>
        <isContainedBy>http://hdl.handle.net/12345/AAAA</isContainedBy>
        <isSuppliedBy encoding="ens:dataProvider" label="data provider">
           Schweizerisches Sozialarchiv</isSuppliedBy>
        <isRepresentedBy label="persistent identifier digital object">
            <rights encoding="ens:right" label="copyright">http://www.
               europeana.eu/rights/rr-f/</rights>
            <language encoding="dc:language" label="language digital content"
                normalised="gsw">gsw</language>
            <persistentID>http://hdl.handle.net/12345/DR</persistentID>
            <localID>DR1</localID>
            <type>SOUND</type>
        </isRepresentedBy>
        <landingPage encoding="ens:landingPage" label="landingPage" localID="
           LP">http://hdl.handle.net/12345/LP1</landingPage>
        <thumbnail encoding="ens:object" label="thumbnail" localID="TH">http
           ://hdl.handle.net/12345/TH1</thumbnail>
        <metadataLanguage encoding="dc:language" label="language metadata"
           normalised="deu">ger</metadataLanguage>
        <title cataloguing="spectrum:title" encoding="lido:titleSet" label="
           title" language="deu">Versammlung der Z rcher Jugendbewegung vom
           1. Juni 1980</title>
        <date cataloguing="spectrum:object production date" encoding="lido:
           eventSet > lido:eventDate" label="object production date"
           normalised="1980-06-01" script="">1980-06-01</date>
        <provenance cataloguing="spectrum:owner" encoding="lido:eventSet >
           lido:eventActor" label="owner">Vollversammlungen Jugendbewegung
           Z rich</provenance>
        <domain>
            <visualUnit>
                <descriptionLevel encoding="lido:recordType" label="level of
                    description" normalised="item">item</descriptionLevel>
                <objectName cataloguing="spectrum:object name" encoding="
                   lido:objectWorkType" label="object name">Ton</objectName>
                <technicalAttribute cataloguing="spectrum:technical
                   attribute" encoding="lido:objectDescriptionSet" label="
                   technical attribute">intakt</technicalAttribute>
            </visualUnit>
        </domain>
    </descriptiveUnit>
</hopeEntity>
```

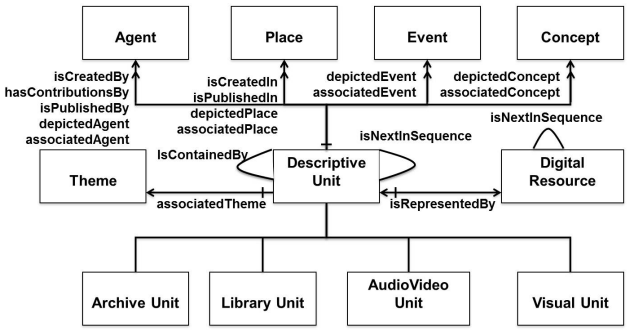

Figure 7 HOPE common metadata model: main entities and relationships

fields whose semantics are defined by the LIDO encoding and the SPECTRUM cataloguing scheme.

Information space. The aggregative information space is populated by collecting and converting metadata records from HOPE content providers and curated by HOPE data curators, who can edit/correct metadata records and tag objects in order to : (*i*) classify them, based on a vocabulary of historical themes (defined as part of the HOPE data model), or (*ii*) establish which social networks they should be sent to, based on a list of social networks. The information space is searchable and browsable by end-users from the project web portal (the Social History Portal (HOPE, 2013)) and made available to Europeana and other interested service consumers via OAI-PMH APIs.

4.3 D-NET and HOPE ADI: two-phase approach

As pointed out in (Haslhofer and Klas, 2010), the use of crosswalks (or mappings) solves structural and semantic heterogeneities of metadata records, enabling the population of homogeneous information spaces where curators and automatic services can operate on. In the case of multiple input data sources, the typical approach is that of defining a common metadata format and establishing mappings from each input format to the common one. In the case of HOPE, this process was complicated by the high degree of heterogeneity. As described above, since the objects and metadata records collected from the content providers may belong to sub-communities of the overall ADI, the HOPE common metadata model tends to abstract over all of such communities and therefore the mapping from data source data models into the common data model is not straightforward. For those reasons, instead of adopting a "classic" crosswalk from one input metadata format to one target metadata format, HOPE ADIs adopts a "two-phase approach". The first phase solves intra-profile structural and semantic heterogeneities, while the second phase solves inter-profile heterogeneities. The first phase is realised by mapping the metadata records of all data sources of the same profile onto metadata records conforming to a given standard data model for such profile; i.e. MARCXML (library), EAD (archive), EN 15907 (audio video), and LIDO (visual). The second phase is accomplished by providing stable mappings from such standard metadata records onto records of the HOPE data model. The approach brings two main benefits: it is easier for data source managers to map their formats into a standard format in their community (in some cases they are adopting the very same standards); and the ADI can export data source content through standard formats without further data processing. Data source managers are guided in the construction of such mappings by ADI data managers, who are in charge of implementing them in the ADI. As previously indicated, data managers can possibly re-use/refine mappings previously defined for similar data sources.

On the other hand, the adoption of standards can be a drawback for data richness in cases where the input format and the common format are richer than the adopted standard. To avoid that loss of richness, the standard profiles adopted in HOPE have been enriched with ad-hoc fields, that providers can use to channel information not regarded by the standard format to the HOPE common format. For example, multilingual descriptions may be lost when mapping onto MARCXML, even if the common format can represent them. Therefore, the library MARCXML profile has been integrated with standard $xml:lang$ attributes, which can be used by provider to map multilingual descriptions in the same MARCXML metadata record.

The crosswalks implemented in the HOPE ADI are described by the aggregation workflow depicted in Figure 8. The workflow describes the two-phase crosswalk required to (*i*) collect the XML records of one data source, unpackage and transform them into the relative HOPE profile records, and (*ii*) to unpackage, transform and clean such records to generate XML records compliant with the HOPE data model. A further crosswalk transforms HOPE XML records into records compliant to the Europeana Data Model (EDM) (Isaac, 2011), which can be accessed via OAI-PMH by Europeana.

D-NET services from different functional areas are configured and combined by data managers to realise the flow, which is automatically orchestrated by the system.

Data mediation services have been configured to handle the (de-)registration and management of a variable set of data sources belonging to different content providers (organisations). Each data source is associated to one of the four HOPE profiles. Services can collect UTF-8 encoded XML metadata records via OAI-PMH, FTP, SFTP, HTTP and HTTPS. Thanks to the flexibility of Data Conversion services, the only requirements for HOPE data sources is to declare the XML schema of the records it will provide and make sure the records will contain one stable identifier. The guarantee of stable identifiers allows for the creation of "stateless" identifiers to distinguish records in the aggregated information space; e.g. *datasource_stableID*.

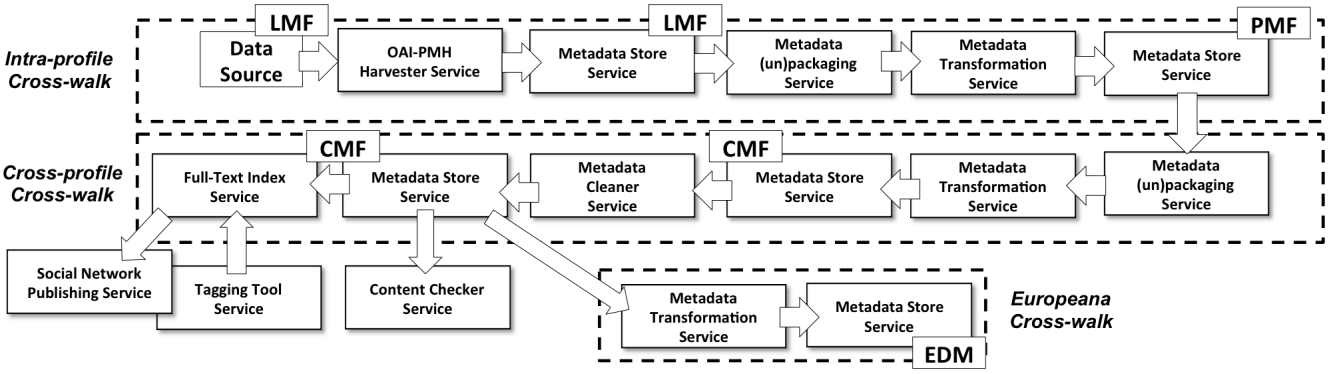

Figure 8 The HOPE aggregation workflows

This means, for example, that when XML records are collected again from the same data sources, possibly with updated information, their corresponding records in the ADI will be properly updated. Not only, all the relationships between such records and other records, will still hold.

Data Conversion services have been combined and customised to realise the "two-phase transformation" and to deliver records to Europeana. In particular, the HOPE ADI contains one running instance of the Harvesting Service, Transformator Service, Cleaner Service, and (Un)Packaging Service, while MDStore Service and Index Service are replicated to ensure robustness and availability. The crosswalks for one data source are implemented by pipe-lining calls to such services parametrized with the proper transformation mappings, cleaning rules and unpackaging modules, defined following the advice of the relative data source manager. The Cleaner Service is applied only in the last phase. It has been customised to clean the values in the metadata fields for languages, countries, level of description, scripts, type of resources, instantiation types, rights, and spatial coverages according to the following standard controlled vocabularies: ISO 639-3 language codes, ISO 3166-1 country codes, ISO 15924 script codes. For resource types and rights, the corresponding Europeana vocabulary has been adopted. Idiosyncratic vocabularies have been generated for level of description and instantiation types. Synonyms have been added based on the values used in the original records. Records which were successfully cleaned are delivered to the index, for use of the Social History Portal, and transformed into EDM records to be OAI-PMH harvested by Europeana. Records which could not be cleaned are marked as invalid and made available for inspection via a dedicated user interface. Data managers can verify which value is invalid and then refine the affected thesaurus accordingly.

As mentioned above, the HOPE information space is handled in the form of XML files, stored in Metadata Store Services and accessible via Full-text Index Services. As shown in Figure 8, the results of harvesting – Local Metadata Format (LMF) – and transformation – Profile Metadata Format (PMF) and Common Metadata Format (CMF) – are stored in MDStore services, in order to make it possible to repeat one phase of the crosswalk without harvesting again the records from the data source or re-applying the previous phase. Data curation and enrichment services, i.e. the Content Checker Service and the Record Tagging Service, have also been deployed. These allow data curators and data source managers to: (*i*) search and browse for records in the information space in order to check the correct implementation of the crosswalks, identify records with insufficient information, check the effectiveness of the cleansing phase, and identify records that need to be updated; and (*ii*) create new virtual, cross-data source collections by tagging records with historical themes or social network publishing tags, e.g. objects tagged with "YouTube" are automatically exported to that social site.

Data provision services export the information space via standard SRW/SRU APIs (REST and Web Services). EDM records produced in the last transformation step are published via OAI-PMH. Social Network Publishing Services have also been deployed to react based on the aforementioned tagging actions.

4.4 HOPE ADI: the accomplishments

The HOPE ADI has today collected about 900,000 DescriptiveUnit metadata records from 14 content providers, each record describing one or more Digital Resources, for a total of about 1,600,000. Input records have been converted into the HOPE common data model, and then delivered to Europeana as XML records in EDM format. HOPE digital objects and metadata records are available from the Social History Portal (HOPE, 2013) and Europeana. The real challenge for data curators and data source managers is to exploit the ADI to maintain a uniform information space of interconnected objects despite of their heterogeneity, by continuously refining the mappings, the vocabularies, the original data source content, and classification tags.

After the end of the project, the International Association of Labour History Institutions (IAHLI) foundation will be in charge of the sustainability of the HOPE ADI by adopting an economy of scale approach involving all partners and future institutions willing to join. From a technical and operational point of view, the International Institute of Social History in Amsterdam has been trained by ISTI-CNR to administrate, operate, and monitor the HOPE ADI and provide support to content providers willing to register their data sources or refine their mappings. In the future, further D-NET functionality will be added to disambiguate and enrich the information space, namely De-duplication Services and Metadata Editor Services.

5 Conclusion

In this paper, we showcased the need for aggregative data infrastructures (ADIs) in the Cultural Heritage (CH) domain and described the important role that enabling software for ADIs can play in the lifetime of such systems. If ADI sustainability issues are a common problem in the realisation of such systems, i.e. due to evolving content, functional and architectural requirements, in the Cultural Heritage domain content interoperability also become particularly challenging, due to the heterogeneity of data models and granularity of representation. In particular, we claimed that the realisation of ADIs with a from scratch approach is not sustainable since the resulting software typically lacks the re-usability, scalability and flexibility features required to cope with ADI evolution in affordable way. On this regard, we presented the D-NET Software Toolkit, a service-oriented framework specifically designed to support developers of ADIs to address the above issues. D-NET adopts a general-purpose and loosely-components approach, providing ready-to-use services that can be configured, extended and composed in workflows to meet the specific community's needs. We demonstrated the effectiveness of D-NET in the CH domain by describing how it has been adopted in the context of the HOPE project for the realisation of an ADI implementing a two-phase approach to metadata record conversion.

6 Acknowledgements

This work is partly funded by the *Heritage of the People's Europe* FP7 EU eContentplus, Best Practice Networks Project: Grant Agreement N. 250549. Its completion would have not been possible without the precious cooperation of all partners of the Project Consortium (HOPE, 2011).

References

Getaneh Alemu, Brett Stevens, and Penny Ross. Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist approach. *New Library World*, 113(1/2):1/2, 2012.

Giuseppe Amato, Franca Debole, Carol Peters, and Pasquale Savino. Multimatch: Multilingual/multimedia access to cultural heritage. In Maristella Agosti, Floriana Esposito, and Costantino Thanos, editors, *IRCDL*, pages 162–165. DELOS: an Association for Digital Libraries / Department of Information Engineering of the University of Padua, 2009.

Apache.org. Apache solr. `http://lucene.apache.org/solr/`, 2011.

Michele Artini, Alessia Bardi, Federico Biagini, Franca Debole, Sandro Bruzzo, Paolo Manghi, Marko Mikulicic, Pasquale Savino, and Franco Zoppi. Data interoperability and curation: The european film gateway experience. In Maristella Agosti, Floriana Esposito, Stefano Ferilli, and Nicola Ferro, editors, *Digital Libraries and Archives*, volume 354 of *Communications in Computer and Information Science*, pages 33–44. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-35833-3. doi: 10.1007/978-3-642-35834-0_6. URL `http://dx.doi.org/10.1007/978-3-642-35834-0_6`.

T. Blanke and M. Hedges. Providing linked-up access to cultural heritage data. *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage. Aarhus, Denmark*, 2008.

Tobias Blanke. From tools and services to e-infrastructure for the arts and humanities. In Simon C. Lin and Eric Yen, editors, *Production Grids in Asia*, pages 117–127. Springer US, 2010. ISBN 978-1-4419-0046-3.

L. Candela, P. Castelli D., Manghi, and P. Pagano. *Recent Developments in the Design, Construction, and Evaluation of Digital Libraries: Case Studies*, chapter Infrastructure-Based Research Digital Libraries, pages 1–17. IGI Global, January 2013. doi: 10.4018/978-1-4666-2991-2.ch001.

Carl Lagoze and Herbert Van de Sompel. The making of the open archives initiative protocol for metadata harvesting. *Library Hi Tech*, 21(2):118 – 128, 2003.

Donatella Castelli, Paolo Manghi, and Costantino Thanos. A vision towards scientific communication infrastructures. *International Journal on Digital Libraries*, pages 1–15, 2013. ISSN 1432-5012. doi: 10.1007/s00799-013-0106-7. URL `http://dx.doi.org/10.1007/s00799-013-0106-7`.

Tom Cramer and Katherine Kott. Designing and implementing second generation digital preservation services: A scalable model for the stanford digital repository. *D-Lib Magazine*, 16(9/10), 2010.

Rita Cucchiara, Costantino Grana, Daniele Borghesani, Maristella Agosti, and AndrewD. Bagdanov. Multimedia for cultural heritage: Key issues. In Costantino Grana and Rita Cucchiara, editors, *Multimedia for Cultural Heritage*, volume 247 of *Communications in Computer and Information Science*, pages 206–216. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-27977-5. doi: 10.1007/978-3-642-27978-2_18.

D-NET Lab. D-net software toolkit. `http://www.d-net.research-infrastructures.eu`, 2009.

Christos Papatheodorou Costis Dallas Panos Constantopoulos Dimitris Gavrilis, Stavros Angelis. Preservation aspects of a curation-oriented thematic aggregator. In *Proceedings of the 10th International Conference on Preservation of Digital Objects, iPRES2013*, pages 246–251, September 2013.

Martin Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Mag.*, 24(3):75–92, September 2003. ISSN 0738-4602. URL `http://dl.acm.org/citation.cfm?id=958671.958678`.

DRIVER. The digital repository infrastructure vision for european research. `http://search.driver.research-infrastructures.eu`, 2007.

Dublin Core Metadata Initiative. Dublin Core Metadata Initiative. `http://dublincore.org`, 2008.

EFG. The european film gateway. `http://www.europeanfilmgateway.eu`, 2010.

ESPAS. Near earth space data infrastructure for e-science. `http://www.espas-fp7.eu/`, 2012.

European Committee for Standardization. En 15907 film identification - enhancing interoperability of metadata - element sets and structures. European Standard ICS 35.240.30; 97.195, European Committee for Standardization, July 2010.

Europeana Foundation. Europeana: Connecting cultural heritage. `http://www.europeana.eu`, 2009.

Europeana Foundation. Europeana semantic elements specification. `http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57`, February 2012.

exist Solutions. The eXistdb. `http://exist-db.org/`, 2012.

Martin Feijen, Wolfram Horstmann, Paolo Manghi, Mary Robinson, and Rosemary Russell. DRIVER: Building the Network for Accessing Digital Repositories across Europe. *Ariadne*, 53, 2007. ISSN 1361-3200.

Katrina Fenlon, Miles Efron, and Peter Organisciak. Tooling the aggregator's workbench: Metadata visualization through statistical text analysis. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012. ISSN 1550-8390. doi: 10.1002/meet.14504901161.

Panorea Gaitanou, Lina Bountouri, and Manolis Gergatsoulis. Automatic generation of crosswalks through cidoc crm. In JuanManuel Dodero, Manuel Palomo-Duarte, and Pythagoras Karampiperis, editors, *Metadata and Semantics Research*, volume 343 of *Communications in Computer and Information Science*, pages 264–275. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-35232-4. doi: 10.1007/978-3-642-35233-1_26.

B Haslhofer and W. Klas. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2), 2010.

Bernhard Haslhofer, Elaheh Momeni, Manuel Gay, and Rainer Simon. Augmenting europeana content with linked data resources. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 40:1–40:3, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0014-8. doi: 10.1145/1839707.1839757.

HOPE. Heritage of the People's Europe. `http://www.peoplesheritage.eu/`, 2011.

HOPE. Social History Portal. `http://www.socialhistoryportal.org/search-collections`, 2013.

Jane Hunter and Carl Lagoze. Combining rdf and xml schemas to enhance interoperability between metadata application profiles. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 457–466, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.372100.

Eero Hyvönen. Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1):1–159, 2012. doi: 10.2200/S00452ED1V01Y201210WBE003. URL `http://www.morganclaypool.com/doi/abs/10.2200/S00452ED1V01Y201210WBE003`.

IAHLI. International association of labour history institutions. `"http://www.ialhi.org/"`.

ICOM International Committee for Documentation. Lightweight Information Describing Objects. http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/lido-technical/specification/, November 2010.

Institute for System Programming RAS. Sedna naative xml database system. http://www.sedna.org/, 2011.

International DOI Foundation, editor. *DOI Handbook*. 2012. doi: 10.1000/182.

Antoine Isaac. Europeana data model primer. http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5, October 2011.

Antoine Isaac, Robina Clayphan, and Bernhard Haslhofer. Europeana: Moving to linked open data. *Information Standards Quarterly*, 24(2?3), September 2012.

KEEP Consortium. Keeping emulation environments portable. http://www.keepproject.eu/ezpub2/index.php, 2009.

Petr Knoth and Zdenek Zdrahal. Core: three access levels to underpin open access. *D-Lib Magazine*, 18 (11/12), 2012.

Carl Lagoze and Herbert Van de Sompel. The OAI Protocol for Object Reuse and Exchange. http://www.openarchives.org/ore/, 2006.

Bert Lemmens, Joris Janssens, Ruth Van Dyck, Alessia Bardi, Paolo Manghi, Eric Beving, Kathryn Máthé, Katalin Dobó, and Armin Straube. The common hope metadata structure, including the harmonisation specifications. deliverable 2.2. Technical report, HOPE - Heritage of the People's Europe, 2011. URL http://www.peoplesheritage.eu/pdf/D2_2_MetadataStructure.pdf.

Yuan Li and Meghan Banach. Institutional repositories and digital preservation: Assessing current practices at research libraries. *D-Lib Magazine*, 17(5/6), 2011. URL http://www.dlib.org/dlib/may11/yuanli/05yuanli.html.

Library of Congress. Encoded archival description. http://www.loc.gov/ead/, 2002.

Library of Congress. MARC Standards Web Page. http://www.loc.gov/marc/, September 2005.

Claudia Loebbecke and Manfred Thaller. Digitization as an it response to the preservation of europe's cultural heritage. In Andrea Carugati and Cecilia Rossignoli, editors, *Emerging Themes in Information Systems and Organization Studies*, pages 359–372. Physica-Verlag HD, 2011. ISBN 978-3-7908-2739-2.

Paolo Manghi and Marko Mikulicic. Pace: A general-purpose tool for authority control. In Elena Garcia-Barriocanal, Zeynel Cebeci, Mehmet C. Okur, and Aydin Ozturk, editors, *Metadata and Semantic Research*, volume 240 of *Communications in Computer and Information Science*, pages 80–92. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-24731-6. 10.1007/978-3-642-24731-6_8.

Paolo Manghi, Leonardo Candela, and Pasquale Pagano. Interoperability Patterns in Digital Library Systems Federations. In *Proceedings of the Second DL.org Workshop on Making Digital Libraries Interoperable: Challenges and Approaches, in conjunction with ECDL 2010*, Glasgow, Scotland (UK), September 2010a. ISTI-CNR.

Paolo Manghi, Marko Mikulicic, Leonardo Candela, Michele Artini, and Alessia Bardi. General-Purpose Digital Library Content Laboratory Systems. In *Proceedings of the 14th European Conference on Digital Libraries*, Glasgow, UK, September 2010b.

Paolo Manghi, Marko Mikulicic, Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. *D-Lib Magazine*, 16(3/4), March/April 2010c. ISSN 1082-9873. doi: doi:10.1045/march2010-manghi.

Cezary Mazurek, Marcin Mielnicki, Aleksandra Nowak, Maciej Stroinski, Marcin Werla, and Jan Weglarz. Architecture for aggregation, processing and provisioning of data from heterogeneous scientific information services. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 529–546. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-35646-9. doi: 10.1007/978-3-642-35647-6_32.

Jerome McDonough. Structural metadata and the social limitation of interoperability: A sociotechnical view of xml and digital library standards development. In *In Proceedings of Balisage: The Markup Conference*, volume 1 of *Balisage Series on Markup Technologies*, August 2008.

MICHAEL Culture Association. Multilingual inventory for cultural heritage in europe. http://www.michael-culture.org/, 2005.

MongoDB. Mongodb. http://www.mongodb.org, 2012.

OpenAIRE. Open Access Infrastructure for Research in Europe. http://www.openaire.eu, 2010.

Christos Papatheodorou. On cultural heritage metadata. *International Journal of Metadata, Semantics and Ontologies*, 7(3):157–161, 01 2012. doi: 10.1504/IJMSO.2012.050184.

Christos Papatheodorou, Costis J. Dallas, Christian Ertmann-Christiansen, Kate Fernie, Dimitris Gavrilis, Maria Emilia Masci, Panos Constantopoulos, and Stavros Angelis. A new architecture and approach to asset representation for europeana aggregation: The carare way. In Elena García Barriocanal, Zeynel Cebeci, Mehmet C. Okur, and Aydin Öztürk, editors, *MTSR*, volume 240 of *Communications in Computer and Information Science*, pages 412–423. Springer, 2011. ISBN 978-3-642-24730-9. URL `http://dblp.uni-trier.de/db/conf/mtsr/mtsr2011.html#PapatheodorouDEFGMCA11`.

Diogo Reis, Nuno Freire, Hugo Manguinhas, and Gilberto Pedrosa. Repox – a framework for metadata interchange. In Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, editors, *Research and Advanced Technology for Digital Libraries*, volume 5714 of *Lecture Notes in Computer Science*, pages 479–480. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-04345-1.

Tomasz Rosiek, Wojtek Sylwestrzak, Aleksander Nowinski, and Marek Niezgódka. Infrastructural approach to modern digital library and repository management systems. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 111–128. Springer, 2013. ISBN 978-3-642-35646-9.

Thomais Stasinopoulou, Lina Bountouri, Constantia Kakali, Irene Lourdi, Christos Papatheodorou, Martin Doerr, and Manolis Gergatsoulis. Ontology-based metadata integration in the cultural heritage domain. In DionHoe-Lian Goh, TruHoang Cao, IngeborgTorvik Sølvberg, and Edie Rasmussen, editors, *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822 of *Lecture Notes in Computer Science*, pages 165–175. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-77093-0. doi: 10.1007/978-3-540-77094-7_25.

The Apache Software Foundation. Apache HBASE. `http://hbase.apache.org/`, 2013.

UNESCO. Convention concerning the protection of the world cultural and natural heritage, article 1. `http://whc.unesco.org/en/conventiontext`, 1972.

University of Michigan. Digital library extension service. `http://www.dlxs.org/`, 2006.

Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. Clarin: Common language resources and technology infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Shenghui Wang, Antoine Isaac, Stefan Schlobach, Lourens van der Meij, and Balthasar Schopman. Instance-based semantic interoperability in the cultural heritage. *Semantic Web*, 3(1):45–64, 01 2012. doi: 10.3233/SW-2012-0045. URL `http://dx.doi.org/10.3233/SW-2012-0045`.