The Heritage of the People's Europe Project: An Aggregative Data Infrastructure for Cultural Heritage

Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, Marko Mikulicic, and Franco Zoppi

Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Via Moruzzi 1, 56124 Pisa, Italy
name.surname@isti.cnr.it

Abstract. HOPE (Heritage of the People's Europe) is a "Best Practice Network" for archives, libraries, museums and institutions operating in the fields of social and union history. The project provides unified access to materials about the European social and labour history from the 18th to 21st centuries. HOPE proposes guidelines and tools for the management, aggregation, harmonisation, curation and provision of digital Cultural Heritage (CH) metadata and digital objects. Moreover, it offers to institutions joining the HOPE network an operational Aggregative Data Infrastructure (ADI) for the collection, aggregation and access of metadata records from CH content providers. The HOPE ADI is realized using and extending the D-NET Software Toolkit, an enabling framework for data infrastructures.

Keywords: cultural heritage, aggregation, metadata records, mapping, service-oriented architectures, data infrastructures, D-NET.

1 Requirements of the HOPE Community

The Heritage of the People's Europe project (HOPE) provides a unified entry point 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. The community is willing to share an aggregated information space and deliver digital cultural objects, including videos (e.g. documentaries on labour movements), pictures (e.g. photos from Gulags), drawings (e.g. posters from the "Commune de Paris"), and archival documents (e.g. newspapers of migrants), in turn described by highly heterogeneous metadata representations. The goal is to group and interlink such objects in order to establish opportunities for a new cross-country, cross-institution social history background.

To this aim, the HOPE community requires an Aggregative Data Infrastructure (ADI) [6] able to handle a varying number of content providers, which in turn deliver several data sources, each dedicated to store metadata records and files relative to

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 77–80, 2014.

different object typologies. Indeed, as it often happens in the Cultural Heritage (CH) domain, content providers may deliver data sources whose objects belong to diverse sub-communities (in HOPE 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 may store objects of different formats (e.g. images, videos, audio, text material) and different descriptive data models and relative metadata formats. For example, librarians and archivists typically model their digital objects according to different data models and schemata (e.g. MARC[1] for libraries, and EAD[2] for archives), but each of them may have a variety of ways to describe their objects. Furthermore, data sources may export their content via several standard protocols, such as OAI-PMH, FTP, etc. At the end of the project, a total of about 900,000 metadata records will be aggregated, describing around 3,000,000 files in the CH domain. HOPE digital objects will be available from the IAHLI[3] portal and delivered to Europeana[4] as XML records in EDM format.

2 The HOPE Infrastructure

Institutions joining the HOPE network benefit of an advanced distributed ADI which enables them to enhance the quality and the visibility of the digital cultural objects they preserve. Moreover, the project also delivers a Shared Object Repository dealing with the management of digital files for HOPE partners who cannot afford the cost of a local object file store. It allows institutions to deposit their files and it automatically applies conversion algorithms to create files in standard formats and with sizes suitable for web dissemination.

The HOPE ADI. The HOPE ADI is implemented using the D-NET [3][4] Software Toolkit. D-NET is an open source, general-purpose software conceived to enable the realization and operation of ADIs and to facilitate their evolution in time. D-NET implements a service-oriented framework based on standards, where ADIs can be constructed in a LEGO-like approach, by selecting, customizing, and properly combining D-NET services. The resulting ADIs are systems that can be re-customized, extended (e.g. new services can be integrated), and scale (e.g. storage and index replicas can be maintained and deployed on remote nodes to tackle multiple concurrent accesses or very-large data size) at run-time. D-NET offers a rich and expandable set of services targeting data collection, processing, storage, indexing, curation, and provision aspects. In the HOPE implementation, the D-NET toolkit is extended to include new services such as the Record Tagging and the Social Network Publishing and to adopt a "two-phase approach" to metadata records conversion. The ADI allows for the construction of an aggregated information space, populated by collecting (via OAI-PMH, HTTP, FTP) records from HOPE content providers and converting them

[1] http://www.loc.gov/marc/
[2] http://www.loc.gov/ead/
[3] International Association of Labour History Institutions, http://www.ialhi.org
[4] Europeana, http://www.europeana.eu

into a common HOPE format. Moreover, HOPE data curators can edit the aggregated records or tag them, in order to: (i) classify them, based on a vocabulary of historical themes defined by the consortium, or (ii) establish which social networks they should be sent to, based on a list of possible targets. Finally, the ADI makes the information space searchable and browsable by end-users from the project web portal and delivered to Europeana and other service consumers via OAI-PMH APIs.

The HOPE Common Data Model and XML Schema. The HOPE community comprises four "data provider profiles", namely library, archive, visual, audio video. Based on these, the project agreed on a common metadata model and its corresponding XML schema. 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[5] for audio video, and LIDO[6] for visual. The model includes seven classes of interrelated entities: Agent, Place, Event, Concept, Digital Resource, Theme, and Descriptive Unit (DU). DUs represent digital objects and include information about the real world object. According to the profiles, the DU class has four subclasses containing properties that are peculiar to one specific profile. Cross-domain properties are instead defined in the DU super class. DUs are related with each other via containment and sequential relationships so that it is possible to represents hierarchies of objects. A digital resource contains technical information about a digital representation of the object and is linked to the corresponding DU. Digital resources related to the same descriptive units can express sequential relationships, thus establishing a "reading path". Agents, places, concepts, events, and themes contextualize the object and are linked to DUs via relationships whose names describe the semantics of the association.

A "Two-Phase Approach" to Metadata Convertion. As pointed out by Haslhofer and Klas in [5], the use of mappings from each input format to the common format solves structural and semantic heterogeneities of metadata records, thus enabling the realization of homogeneous information spaces. In the case of HOPE, this process was complicated by the high degree of heterogeneity of input data sources: since the objects and metadata records collected from the content providers may belong to subcommunities of the overall ADI, the HOPE model tends to abstract over all of such communities and therefore the mapping from source models into the common model is not straightforward. For those reasons, the HOPE ADI implements 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 realized 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 mappings from such formats to the HOPE format.

[5] http://filmstandards.org/fsc/index.php/EN_15907

[6] http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/

The approach brings two main benefits: it is easier for data source managers to map their formats into a standard format in their community; and the ADI can export data source content through standard formats without further data processing. On the other side, the adoption of standards can be a drawback for data richness in cases where the input format is richer than the adopted standard. For example, multilingual descriptions may be lost when mapping onto MARCXML. Once records are in the common format, their content is harmonized by applying vocabularies established by the consortium and compliant to standards (e.g., ISO country, ISO language). Moreover, curation and enrichment tools are available for data experts in order to: (i) check the quality of aggregated metadata record; (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. Finally, curated records are also transformed into the Europeana Data Model[7] (EDM) to be OAI-PMH harvested by Europeana. Social Network Publishing Services have also been deployed to react based on the aforementioned tagging actions.

Acknowledgements. This work is partly funded by the HOPE “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 the whole Project Consortium (http://www.peoplesheritage.eu/content/partners.htm).

References

1. HOPE Project, `http://www.peoplesheritage.eu`
2. eContentPlus framework, `http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm`
3. D-NET Software Toolkit, `http://www.d-net.research-infrastructures.eu`
4. Manghi, P., Mikulicic, M., Candela, L., Castelli, D., Pagano, P.: Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. D-Lib Magazine 16(3/4) (2010)
5. Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Comput. Surv. 42(2), 7:1–7:37 (2010)
6. Bardi, A., Manghi, P., Zoppi, F.: Aggregative Data Infrastructures for the Cultural Heritage. In: Dodero, J.M., Palomo-Duarte, M., Karampiperis, P. (eds.) MTSR 2012. CCIS, vol. 343, pp. 239–251. Springer, Heidelberg (2012)

[7] `http://pro.europeana.eu/edm-documentation`