Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MARCO KLINDT, KILIAN AMRHEIN, ANJA MÜLLER,
WOLFGANG PETERS-KOTTIG

Dealing with all the data – Participating in workflows to transform digital data about cultural heritage objects within a digital long-term preservation infrastructure[1]

[1]to appear in: In Proc. EVA Berlin 2013, Gesellschaft zur Förderung angewandter Informatik e.V., 2013.

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Dealing with all the data – Participating in workflows to transform digital data about cultural heritage objects within a digital long-term preservation infrastructure

Marco Klindt, Kilian Amrhein, Anja Müller, Wolfgang Peters-Kottig
Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)
Email: klindt, amrhein, anja.mueller, peters-kottig @zib.de

Abstract:
Cultural heritage institutions are on the verge of making their artefacts available in digital form. During this transition they are faced with conceptual and technical challenges that have only little overlap with their traditional domains but provide them with a lot of opportunities. We aim at empowering them to deal with some of these challenges by designing workflows attached to the data flow within a digital long-term preservation system. The preservation framework processes data by utilising micro-services. These are tailored to accommodate data transformations that can help institutions making their data available if they choose to participate in the interconnected digital world

Introduction

The purpose and mission of cultural heritage institutions like galleries, libraries, archives, and museums (GLAM) is to acquire, conserve, and research cultural heritage objects and make these accessible to the public and research. A lot of objects from these vast troves of knowledge are now in the process of being documented with the help of information technology. Computers are indispensable tools in keeping tabs on inventories and the management of collections. Additionally the physical artefacts themselves are being brought into the digital domain by digitisation.
The 2013 Horizon Report [1] identified open-content, that is accessible digital learning resources, as an important immediate imperative memory institutions are facing.
Safekeeping and conserving the physical object is hard enough already but the same has to be done for the digital surrogates as well.
The digital world is, alas, inherently complex even for experts, and preserving the data, while maintaining the connections and meanings of objects and documentation takes a lot of effort. A lack of technical expertise and resources could be compensated by utilising services offered by technical service providers.
Digital assets consisting of descriptive metadata and representations destined for long term preservation require a rather high quality with respect to interoperability and semantics of the metadata to describe various aspects of the artefacts as well as parameters for the acquisition. These requirements lead to complex metadata structures to adequately describe the diversity of scholarly disciplines on the one side and on the other side to huge amounts of binary data, e.g. digital images, or audio-visual files.
The Zuse Institute Berlin, a research institute for applied mathematics and computer science, is establishing the infrastructure and services for digital long term preservation of cultural heritage datasets based on the existing infrastructure providing long term storage of binary data generated by simulations and calculations from the natural sciences' domain. The Servicestelle Digitalisierung (Service Center Digitisation) supports partner GLAM institutions in Berlin with technical expertise and offers a long-term preservation infrastructure for the outcome of digitisation campaigns.

From long term storage to digital preservation

The retention of digital data for longer periods of time must ensure the unaltered recall of binary data that is identical to the original data. The data has to be stored redundantly with two or more

copies. Error detection and correction have to take place to prevent data loss. Hardware and infrastructure has to be maintained and upgraded to be able to read back the submitted information packages on a physical level. Fixity information, i.e. mathematical codes that can be used to prove the information hasn't changed, has to be provided and monitored for each data object. And finally the data has to be migrated to new or upgraded storage media in case a particular medium might get unreadable or becomes obsolete in the future.
Provided the challenges of digital long-term storage are already being taken care of, the task of digital preservation aims to maintain the interpretability and meaning of contents. Semantics of metadata fields and renderability of binary formats are preserved through migration, i.e. the transformation to other, viable file formats and ontologies, controlled vocabularies or thesauri. Metadata for the purpose of digital preservation also include technical identification data and descriptions of the file formats used in the archival system, rights statements governing the use and access of the datasets, unique identification and relations to external datasets. The information in an archival package also needs to be of such quality and richness that the content is independently understandable by the identified designated community.
The tasks involved in the digital preservation process include constantly performing a technology watch to track adoption trends in the use of file and metadata formats, migrating if necessary, and documenting meticulously every event the data might encounter during its lifetime within the system, e.g. physical transfers and transfers of custody, or transformation to different file formats, metadata formats and container formats and taking note of any software version that is used during processing.

Workflows for digital preservation

In digital preservation systems (DPS) the raw data in the form of a data object can be interpreted using the represented information in combination with the description of context (domain, ontologies, dictionaries, grammars and so on) and thus yields an information object that should enable a member of a designated community to reconstruct the retained knowledge (with an intellectual effort).
A digital preservation system is divided into distinct components that handle different phases of the archival data flow of objects. The functional entities are described in ISO 14721:2012 as the Open Archival Information Systems (OAIS) model [2]. The top-level responsibilities are ingest, storage, and access. Data from producers are ingested in the form of *submission information packages* (SIPs). The digital archival storage and preservation management of the representation information is done with *archival information packages* (AIPs), and the access component delivers information objects in the form of *dissemination information packages* (DIPs) to a consumer. The interaction and logic in and between these components is guided by policies, i.e. technical rules chained together into workflows.
We use the open-source digital preservation system Archivematica [3] that implements the OAIS functional model. The system provides services following the micro-service design pattern. A micro-service is an independently executable task that is responsible for delivering a specific outcome or transformation. Micro-services are for example transferring files from one folder to another, creating a specific folder structure, identifying the file type, performing a virus check, comparing data, transforming one data file format to another. Micro-services can be chained together to perform higher-level tasks called jobs. These jobs can be integrated into workflows that represent business logic. Each micro-service runs either successfully or fails. Depending on the results of certain micro-services and user controllable decisions different branches can be followed. The advantage of this approach is the flexibility in respect to possible actions and how goals can be achieved.

As any data must pass through Archivematica anyway in order to being deposited into long term storage and preservation, it seems natural to implement these additional tasks as jobs or workflows within this system.

Working with data

GLAM institutions index their objects in collection management software (CMS) tools that allow them to describe and document artefacts, to create references to actors and events relevant to creation and provenance, and to manage exhibitions, loans and storage. For these purposes preview images in lower resolution or other media to illustrate the artefact in a dataset is often sufficient to work with. Although there might exist high quality digitalisations of the physical objects, these so-called master-files aren't necessarily managed within a CMS but live outside the system either in a separate asset management repository or simply in a file system. To be of use for the purpose of digital preservation both information sources together are needed to describe the physical object reasonably well.
We have identified six main areas where a digital preservation system and infrastructure could help GLAM institutions handling their data:

Data consolidation: Gather data belonging to one intellectual entity from separate data sources.
Data consistency: Check data for compliance regarding file formats, specification, or controlled vocabularies.
Data enrichment: Embed relevant additional data into data sets from external data sources.
Data presentation: Export a subset of the data as views for reports or web sites.
Data reference and cross-reference: Supply resolvable persistent identifiers for data sets and check for validity regarding external references.
Data delivery: Provide services for data harvesting or download conforming to standardised protocols.

These areas are discussed in more detail in the following sections. As an example we step through the different areas with the use case of a digitised self-portrait of Karl Hagemeister, a painter from the Berlin Secession artist movement, and its metadata from our project partner Bröhan Museum, Berlin [4]. This exemplary workflow is illustrated in figure **1**.

Data consolidation

As the description of a digital object may contain data from separate data sources rules have to be defined that govern the creation of an aggregation containing data belonging together. If the data that comprises a digital object is already in a container (e.g. a folder, a zip archive, or a *metadata encoding transmission standard* (METS) file [5]), that container must only be transformed into an information package that can be ingested into the preservation system. If it is not, rules have to be agreed upon that depend on how an institution manages their data. Some common conventions to link different data files may be a reference to the file inside the descriptive metadata export of the CMS or files are named with the same name but different extensions. If the association is consistently done and unambiguous a workflow can automatically assemble packages from the raw data files.
Before proceeding the completeness of each digital object has to be checked. Incomplete data has to be then reported back to the submitting institution.
Bröhan Museum includes a link to an preview jpeg image in their CMS, exports the metadata mapped into the LIDO metadata standard [6] and supplies the master images in tiff format named exactly like the preview image. The DPS transforms the link to the jpeg representation and checks if a master image of the same name exists. The metadata and the master-files are then linked within the system as seen in figure **1(b)**.

Data consistency

The descriptive metadata, although originating from a single CMS most of the time, may not be consistent within an institution. Reasons for that could be typing errors or it could stem from different researchers or departments who might have different approaches regarding the semantics of a certain metadata field or use different terminology. Data that can be examined for consistency on a technical level are the presence of certain file formats, validity of file formats, the format and content of metadata fields (e.g. character encoding, date formats, a specific text field cannot include a certain word, entry is not from a controlled vocabulary, etc.). In addition to such consistency checks, another simple task should check the existence of data in mandatory fields.
Anything that can be expressed in a formalised, rule-based workflow covering every exception can also be transformed at this stage. Great care has to be taken if the system is allowed to alter the data.
In our example the metadata about Karl's Painting is complete in regard of the agreed mandatory fields: it contains a title, the creator, the creator's vital dates, a description of materials and techniques used in the creation process, a creation date or timespan, measurements, and references to publications describing the art work in the CMS. The vital dates are recorded as strings not adhering to the ISO 8601 date formatting standard but they can still be unambiguously converted to the desired date format. Because there is only one entity participating in the creation process the creation date should lie between the birth and death dates of its creator. In this case it is and processing can continue as shown in figure **1(c)**.

Data enrichment

Additional information on the semantic level can be added automatically to digital objects by using trusted knowledge repositories like authority files, e.g. the Virtual International Authority File (VIAF), an aggregation of a number of authority files for personal and corporate names, classification schemes, e.g. the Library of Congress subject classification, or thesauri, e.g. the Getty Thesaurus of Geographic Names (TGN) or the Art & Architecture Thesaurus (AAT).
If no such reference exists, the system still can heuristically search for a reference. For example given a name and vital dates the system could automatically search for the place of birth, or get an authoritative official name given some variant name of a place and perhaps coordinates.
The name Karl Hagemeister with vital dates matches exactly one person in the VIAF and is referring to entries in the Getty Union List of Artist Names (ULAN) and the German Integrated Authority File (GND). From both sources additional data is integrated by the DPS into the dataset: Hagemeister is male, German, and was born and died in Werder (Havel) (see figure **1(d)**).
Such an enrichment of data needs to be always marked as additional information with both reference system and the preservation system as the source. The enriched data has to be reviewed intellectually by human operators, as even knowledge authorities may contain errors. The enriched data is then stored on tape for long-term retention as shown in figure **1(e)** as well as further processed for presentation purposes.

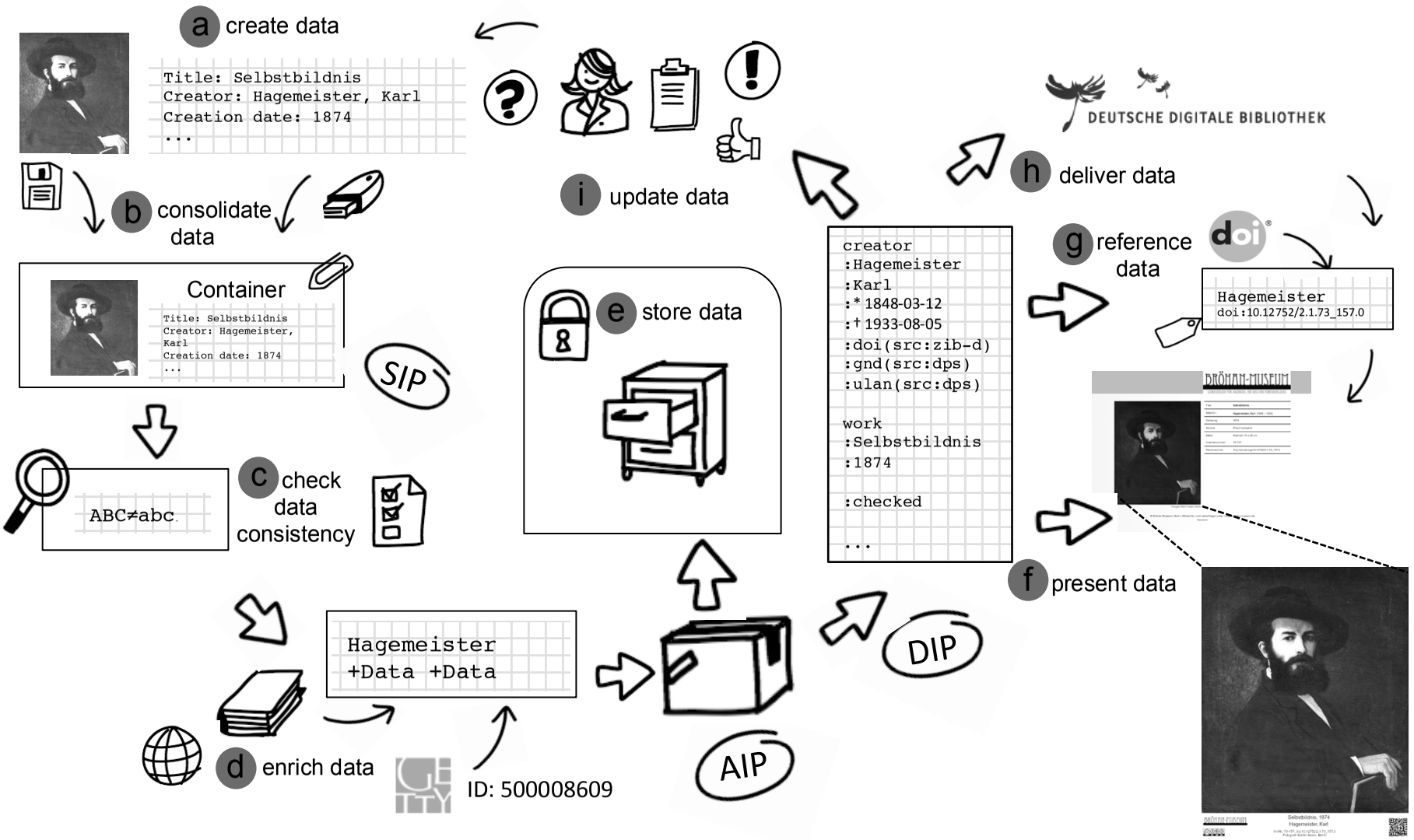

Figure 1: Data workflow

Data presentation

The preservation system holds all data that are necessary to create websites in order to display the digital object comprising metadata and digital representation. This data can be exported as described in the section about data delivery but the preservation system could also provide workflows to generate parts of or a whole web page. The display of metadata as text is trivially embeddable into such a page as well as hidden in micro-data formats for semantic indexing through search engine crawlers. Images on the other hand can be transformed in many ways automatically: Annotations can be added either as machine extractable tag entries in the image files themselves or be imposed on top of an image or rendered in areas outside the depiction of the object accompanied by logos or other pictograms, even as QR codes that can be scanned with handheld devices. Additionally visible and invisible digital watermarks could be added to images, video, or audio files if required by a policy.
In the Hagemeister example selected factual metadata is exported and inserted with a preview image into a static web page. The page follows design guidelines set by the museum. A higher resolution image with an annotation-footer will be shown if the web user clicks on the image. The final result can be seen in figure **1(f)**. This web page can be indexed by search engines and provides the museum with a higher visibility for marketing or research purposes.

Data reference and cross reference

An important part of research publications is the ability to reference ideas and the subjects of research. Regular internet-links in the form of Uniform Resource Locators (URLs) have the disadvantage that they refer to a specific location on a specific machine that has to be always available and must never change for others to access the referenced information. Persistent Identifiers (PIs) are an approach to facilitate the stable access to information through the use of an intermediate resolver service, which separates the data location from the locator. The identifier thus becomes a stable reference for quotation of information.
The DPS also checks if external dependencies are resolvable at ingest time. This does not guarantee access in the long-term but at least provides a report on the status quo.
Our DPS assigns Digital Object Identifiers (DOI). These rely on the global handle system to resolve identifiers to URLs and also embed additional metadata beneficial for the purpose of citation. The publication of a DOI is deferred until a view of the digital object is publicly accessible (see next section).
In our example we register the DOI 10.12752/2.1.73_157.0 to point to the data associated with Karl's self-portrait as shown in figure **1(g)**.

Data delivery

The verified and possibly enriched data packages can be made available through a repository system to authorised parties. The originating institutions could thus import the complete set or parts of the data back into their management systems. The repository can also provide different views to that data, i.e. a subset of metadata in a certain format or a specified smaller version of the master representation.
Discovery and exploration of cultural heritage data sets are important aspects of dissemination. Different portals are able to automatically harvest a subset of metadata and thumbnail or preview images directly from such a repository by means of standardised protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) or SWORD (Simple Web-service Offering Repository Deposit) and predefined metadata schemes.
Selected parts of the metadata about Karl Hagemeister are compiled into a record with embedded references to the locations of preview and thumbnail images and the DOI. The record is then transmitted to the German Digital Library (DDB) [7] for ingest into the portal as seen in figure **1(h)**.

Conclusion

The digital long-term preservation of digitised cultural heritage artefacts requires not only high quality digital representations but also the application of widely used, complex, and heterogeneous metadata schemes. The huge amount of data and complex standards generated by digitisation campaigns can easily exceed the abilities found in smaller memory institutions as it requires additional technical expertise and organisational resources that go beyond the day-to-day work of collecting, describing, exhibiting and keeping safe the physical objects.
A lot of the challenges that arise in transferring, ingesting, storing, and making accessible digital objects have to be addressed in our digital long-term preservation system anyway (figure **1(e)**) and thus provide us with the opportunity to tailor some of the micro-services in a way that empowers content providers to work with their data through the detour of digital preservation in obtaining consistent, standardised, high-quality data and metadata for reuse in their daily workflows (figure **1(i)**). The services offered will not make the museum curators and researchers obsolete but empower them to work with their digital assets more effectively.

Acknowledgements

The digital long-term preservation system described is being implemented at the Servicestelle Digitalisierung at the Zuse Institute Berlin and is supported by the State of Berlin. The dataset describing Hagemeister's painting was provided by the Bröhan Museum, Berlin. The corresponding image was photographed by Martin Adam.

References

[1] Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., Ludgate, H.: *NMC Horizon Report: 2013 Higher Education Edition*. The New Media Consortium, Austin, Texas , 2013.

[2] *Reference Model for an Open Archival Information System* (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1, August 2009.

[3] C. Mumma, P.V. Garderen: *Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation* in Proc. 10th International Conference on Preservation of Digital Objects (iPres 2013). ISBN 978-972-565-493-4 Biblioteca Nacional de Portugal, Lisbon, 2013.

[4] Bröhan-Museum, Berlin: *Hagemeister, Karl: Selbstbildnis*; Bröhan-Museum, Berlin, 2013. http://dx.doi.org/10.12752/2.1.73_157.0

[5] Library of Congress: *Metadata Encoding and Transmission Standard*, http://www.loc.gov/standards/mets/mets-schemadocs.html [accessed 2013-09-01]

[6] ICOM-CIDOC Working Group Data Harvesting and Interchange: *Lightweight Information Describing Objects,* http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf [accessed 2013-09-01]

[7] Deutsche Digitale Bibliothek: Hagemeister, Karl: Selbstbildnis; http://www.deutsche-digitale-bibliothek.de/item/4STYYV6ALWO53SARNW56ZNAYCIVJTCKZ [accessed 2013-09-01]