A Generalised XML Client for the Bavarian State Archives

Markus Schmalzl (Directorate General of State Archives Bavaria, Germany)

Markus Schmalzl is an archivist and works at the Directorate-General of the Bavarian State Archives, previously at the State Archive for Upper Bavaria and the Bavarian Main State Archive at Munich. He is responsible for the transfer of digital data from EDRMS, geodatabases and other collections to archives and for developing interfaces for transfer processes. He is member of committee of the German Conference of State Archive Directors (KLA). He teaches at the Bavarian Archives School, the Bavarian University of Applied Sciences for Public Service and the University of Regensburg.

Schmalzl presented a one-size-fits-all approach for a standardised way of exporting data from government information systems that has been developed in the last few years: a new tool for database archiving.

Databases are widely used by state agencies. Not all the data are important for the long term, but some will need preservation. There is valuable information in high volumes with potential for scientific reuse and other purposes. The State Archives need interfaces for ingesting data from Bavarian government agencies. Data are sometimes volatile and can include Binary Large Object (BLOB) content. Typically, there are no managed archival interfaces for handling this type of information. The State Archives did not want to use emulation but instead applied an approach based on format migration.

Schmalzl's employer wanted to create a standard solution which would work for many systems. The key issue was to choose in which format this information was to be extracted. To close the gap the Bavarian State Archives created a generalised XML extraction tool. It was set up in a test environment October 2021. It is an archival client solution that transfers data in a fully automated way.

The project came to life in 2017 as a specialised client for transferring Bavarian government staff records from the SAP Human Resources (HR) software as PDF files, accompanied by XML metadata. As a user of this client system, you can receive, validate, and save the data and confirm receipt. The interface took a long time to develop. To use the client flexibly for other data sources, the 2017 project was extended to extend use to XML data objects of various schemas that conform to the standards of the different government branches. A mapping tool is the central component. The mapping tool can define how to validate data, whether an appraisal is required and what metadata are presented in the graphical user interface for appraisal. The XML data can be accompanied by text and images.
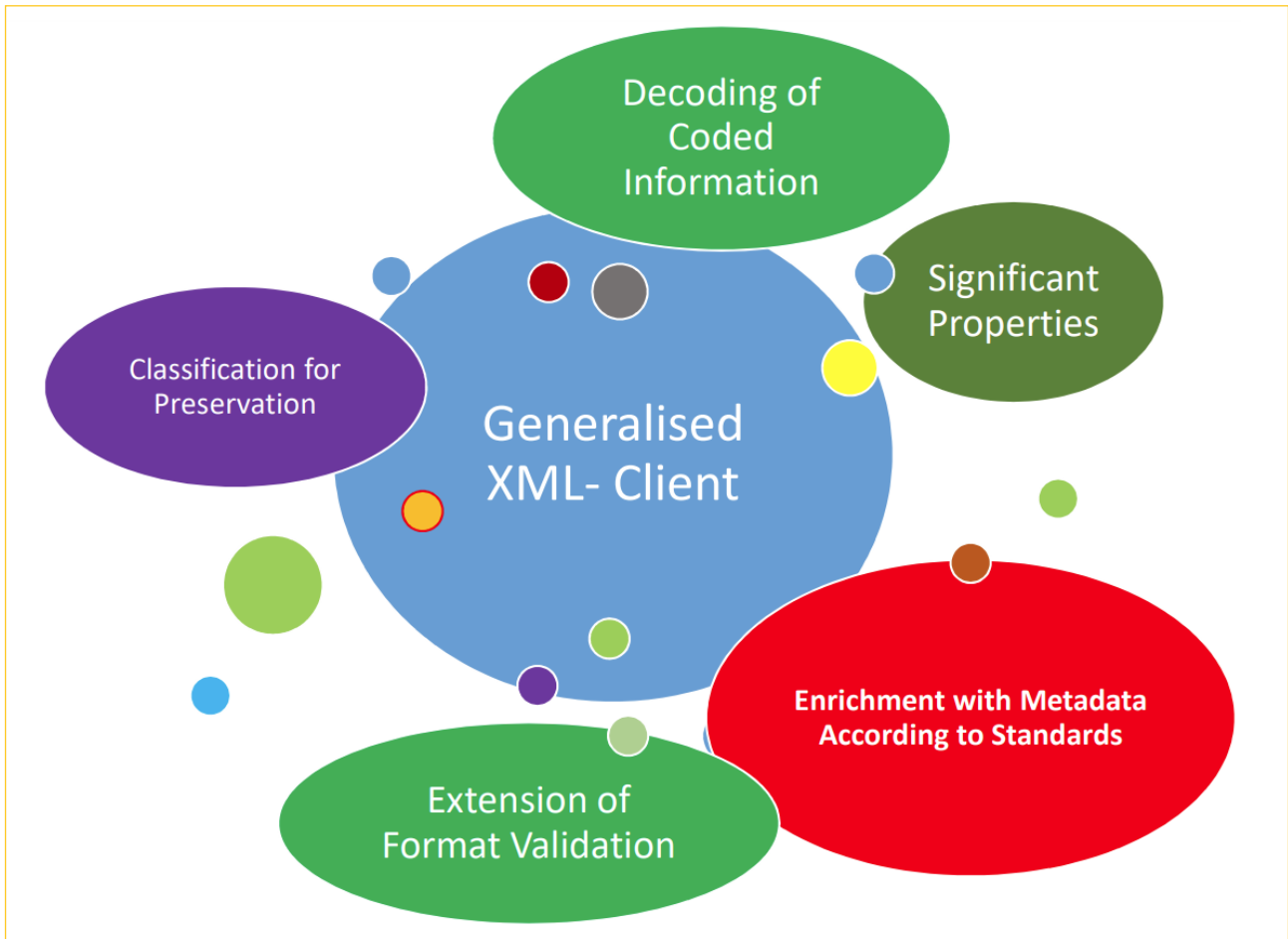
Figure 31: Options for further development of the Generalised XML Client

After configuring the mapping tool, a semi-automated workflow follows. Data are checked continually. Packages are restructured as needed by the mapping tool. The SIP is ingested into archival data storage and structured into AIPs that can be ingested. Attribute descriptions are incorporated into the AIP. Once data has been processed the institution gets a receipt. Hash values are controlled. All archival working steps are recorded: from the reception of the data to appraisal decisions, and finally to successful storage in the archive. The system is automated but flexible. A standardised structure is also maintained for deliveries.

Data are archived in a structured way in an appropriate long term archiving format which facilitates further processing and use. The State Archives do not archive the database management systems themselves. The rather apply this approach to data management systems of government and science.

Further development is planned. Issues relate to significant properties, enrichment with metadata and classification for preservation. Ultimately, the system allows the processing of large data volumes.
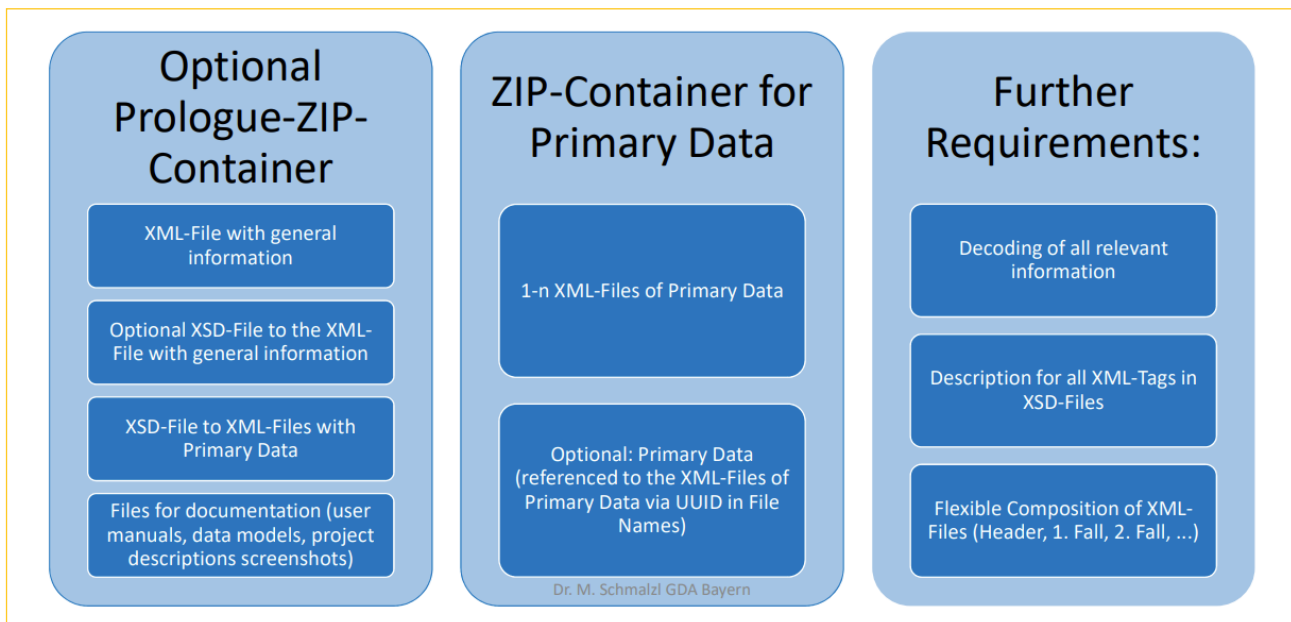
Optional Prologue-ZIP-Container	ZIP-Container for Primary Data	Further Requirements:
XML-File with general information	1-n XML-Files of Primary Data	Decoding of all relevant information
Optional XSD-File to the XML-File with general information	Optional: Primary Data (referenced to the XML-Files of Primary Data via UUID in File Names)	Description for all XML-Tags in XSD-Files
XSD-File to XML-Files with Primary Data		Flexible Composition of XML-Files (Header, 1. Fall, 2. Fall, …)
Files for documentation (user manuals, data models, project descriptions screenshots)	Dr. M. Schmalzl GDA Bayern	

Figure 32: Packaging structure for deliveries through the Generalised XML Client

Questions and discussion

- Will the Bavarian State Archives need a standardised format like SIARD going forward or will they use department standards? The Archives are not using the individual standards of departments currently as they contain partially coded information. It is important to extract certain data for reuse and appraisal. They are open to SIARD but believe this will not guarantee the reusability of the data secured in the databases. The Archives think it might not be the right way to go for the long term but might be convinced of this in the future.
- Kai Naumann explained about coded values in German government databases. Names of court or police stations are coded as values within the XML and the up-to-date key to the code is maintained elsewhere at a special agency at Bremen, another state of the German Federal Republic. Data are well structured, but some tags are not de-coded and this is a challenge when carrying out preservation. Markus confirmed that this is the problem with some German government XML standards.