Xeproc©: A Model-Based Approach towards Document Process Preservation

Thierry Jacquin, Hervé Déjean, Jean-Pierre Chanod

6 chemin de Maupertuis 38240 Meylan France
Firstname.Lastname@xrce.xerox.com

Abstract: Developed in the context of the EU Integrated Project SHAMAN, Xeproc© technology lets one define and design document processes while producing an abstract representation that is independent of the implementation. These representations capture the intent behind the workflow and can be preserved for reuse in future unknown infrastructures. Xeproc© is available under Eclipse Public Licence.

Xeproc© in the context of digital preservation

Xeproc© was developed in the context of the Integrated Project SHAMAN (http://shaman-ip.eu/), co-funded by the European Union within the FP7 Framework. SHAMAN aims at developing a long-term digital preservation framework and tools to analyse, ingest, manage, access and reuse digital objects.

In SHAMAN, Xeproc© use focused on metadata extraction processes [1] operated in the preingest phase. Those processes have been applied to two major types of collections, the Deutsche NationalBibliothek (DNB) collection of electronic PhD theses (available in PDF format) and digitized collections provided by the Göttinger Digitalisierungszentrum (including proceedings and journals). More precisely, the extracted metadata are produced by XML document processing pipelines dedicated to document structure analysis [2, 3, 4]. Eventually processes developed with Xeproc© have been exported and deployed on an iRODS data grid (https://www.irods.org) [5] and the extracted metadata exploited through Chehsire3 (http://www.cheshire3.org/) in support of advance search and navigation [6].

The extracted metadata is stored externally to the document themselves, and can be seen as digital objects to be preserved on their own altogether with persistent document ID to enable preservation management and reuse of the metadata. In this view, the metadata extraction processes belong to the context of production of the metadata. By enabling the preservation of logical descriptions of those processes the Xeproc© methodology provides the ground for documenting the metadata provenance information, i.e. information that documents the history of the Content Information [7], where the content is the metadata in this case.

This will support the long term understanding of the metadata and of the extraction processes and will enable their reconstruction as technology evolves and improves over time.

More specifically, within the context of SHAMAN and digital preservation, Xeproc© models XML pipelines and XML validation checkpoints. These capture the intent behind the workflow irrespective of the implementation at a given point in time. These abstract representations are preserved, so that the Xeproc© models can be seen as independent specifications to be instantiated and deployed over time and as technology evolves. These logical and persistent descriptions, when associated with the accurate components, are interpreted or translated into any SOA orchestration language to produce logically structured documents (typically XML). These make explicit how the source document content is logically and semantically organized.

The Xeproc© technology

Xeproc© technology can be used to build a wide range of applications based on document processing, including transformation, extraction, indexing and navigation. It can be easily integrated with more global business processes and customized to match specific requirements and infrastructures. In the spirit of service-oriented architecture (SOA), Xeproc© embeds references to services and documents and provides loose coupling not only to services but also to data resources, with respect to both their location and format.

Available on Eclipse 3.5.1 under the Eclipse Public License, Xeproc© combines a domain-specific language (DSL), an associated graphic designer and extension APIs (application programming interfaces).

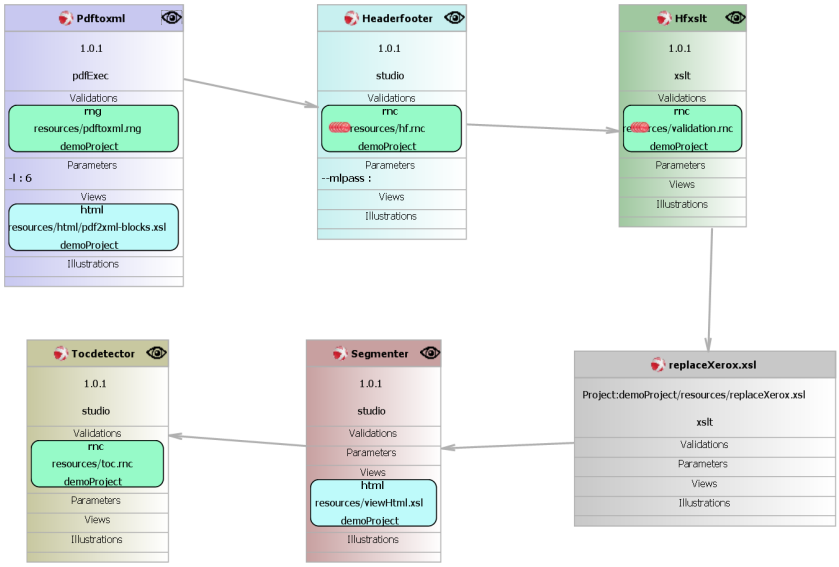

Fig. 1. A Xeproc© template under design.

The Xeproc© DSL: extensible, easy to use and focussed

The Xeproc© Domain-Specific Language (DSL) is used to describe the document process one wants to design. It specifies a chain of processing steps, which may point to components such as document services or project-specific resources. All components take a document as input and generate another document as output.

To take full advantage of Xeproc©, the designer links processing steps with validation resources. While validations are traditionally exploited just before deployment, the Xeproc© Designer is conceived in such a way that they are exploited throughout the design process. Thanks to a continuous monitoring mechanism, validations not only verify but also specify, and lead the design process from the specification to instantiation.

In addition, processing steps can be linked to visualization specifications, highlighting selected outputs. These views, which are captured on demand and throughout the entire monitoring of the process, make it easier to identify and pinpoint errors, undertake corrections or consult the relevant experts.

The Xeproc© DSL is open enough to support any document format, validation syntax and resource location. The Xeproc© DSL is defined by a dedicated XML schema available at http://www.xrce.xerox.com/Xeproc.

The Xeproc© Graphic Designer

The Xeproc© Graphic Designer is a user-friendly Eclipse plug-in editor which allows the user to manipulate abstract representations of objects relevant to the Xeproc© application domain.

The Designer provides an intuitive representation of underlying Xeproc© models and the ability to draw, rearrange and tune document-processing chains. This is achieved by combining project-specific resources (processing components, validations and views) with generic document services organized in a palette. The processing elements are represented as boxes, intermediate documents as arrows and validation constraints and views as icons on boxes.

The Designer was generated from the Xeproc© model using the EMF/GMF (Eclipse Modelling Framework and Graphical Modelling Framework) technologies provided by Eclipse (http://www.eclipse.org/). Model-Driven Architecture methodologies [8] supported by the Object Management Group (http://www.omg.org/) were applied.

Example scenario

A document transformation project will typically create an Eclipse project, share it amongst all the technical partners and initialize it with the reference resources such as documents, requirements and schemas to be validated. The process designer will consider the context and customize the palette of components with those considered useful from a site update. From there (s)he will start the building process and may drag and drop from the component palette or from the project workspace, quickly drawing specific logical and persistent pipelines for document analysis and transformation.

Links: http://www.xrce.xerox.com/Xeproc

Acknowledgements

This work is supported by the Large Scale Integrating Project SHAMAN, co-funded under the EU 7th Framework Programme (http://shaman-ip.eu/).

References

1. Milena Dobreva, Yunhyong Kim, and Seamus Ross. Designing an automated prototype tool for preservation quality metadata extraction for ingest into digital repository. *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies*, 5, 2008.
2. Hervé Déjean, Jean-Luc Meunier, Logical document conversion: combining functional and formal knowledge. ACM Symposium on Document Engineering 2007: 135-143
3. Hervé Déjean, Jean-Luc Meunier: On tables of contents and how to recognize them. IJDAR 12(1): 1-20 (2009) International Journal on Document Analysis and Recognition
4. Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document structure analysis algorithms: a literature survey. In Tapas Kanungo, Elisa H. Barney Smith, Jianying Hu, and Paul B. Kantor, editors, *Document Recognition and Retrieval X*, volume 5010, pages 197–207. SPIE, 2003.
5. iRODS: integrated Rule Oriented Data System. White Paper. Data Intensive Cyber Environments Group. University of North Carolina at Chapel Hill, University of California at San Diego.
6. Sanderson, Robert and Larson, Ray. "Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval", JCDL 2005.
7. 'Reference model for an Open Archival Information System (OAIS).' Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems. 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf
8. http://www.omg.org/mda/